Exploring the involvement of gene regulatory variation in wheat improvement with network analysis

Roos Goessen

Minor thesis: BIF-80324

Department of Bioinformatics

Supervisors at University of Guelph: dr. L. Lukens and dr. E. Raherison

Supervisor at Wageningen University: dr. H. Nijveen

Examiner: prof.dr.ir. D. de Ridder

Date: 08-01-2018

Abstract

Gene regulation differences are often important in crop improvement. This project aimed to explore the involvement of gene regulation differences in wheat improvement using network analysis, an analysis that groups co-expressed genes into modules. Regulation differences can be due to variation in trans-acting factors that affect the expression of downstream genes, or in upstream regions of genes acting on the gene itself (cis-regulatory variation). Many unlinked loci underlying a single trait are difficult to retain in a segregating population due to recombination. We therefore propose that regulation of biological processes involved in plant improvement would favor either single master regulatory loci that control expression of downstream genes or fixed cis-regulatory variation of linked genes, i.e. genes located on the same chromosome.

We tested these hypotheses using RNA-sequencing data from a segregation population derived from heritage and modern spring wheat cultivar cross. We assessed differing biological processes between the parents using differential expression analysis. With network analysis, we aimed to identify co-expressed genes regulated by single regulatory loci and linked genes with cis-regulatory variation that have consistent upregulation or downregulation in one of the parents. We interfered biological processes related to modules.

Among the parents and modules many biological processes related to wheat improvement were identified, such as disease resistances, abiotic stress, flowering time and dwarfing. We identified both modules without linked genes, indicating regulation by single regulatory loci, and modules with linked genes, possibly regulated by cis-acting variation. For the latter, two modules with consistent parental upregulation or downregulation were identified that indicate selection, however no biological processes were inferred. We conclude that single regulatory loci are involved in many cases in the improvement of wheat. No strong evidence indicated that linked genes with cis-regulatory variation are involved in wheat improvement. These results may be shared with other crops.

Table of Contents

Introduction

Wheat (*triticum aestivum*) is one of the most consumed crops worldwide, partially due to its richness in proteins, carbohydrates and minerals. This crop has a complex allohexaploid genome structure, that arose after two polyploidization events (Dubcovsky and Dvorak., 2007). During the first polyploidization event around 0.5 million years ago *Triticum urartu* (AA genome, 7 chromosomes) and an *Aegilops spleltoides* related species (BB genome, 7 chromosomes) together formed *Triticum turgidum* (AABB genome, 14 chromosomes). At the second polyploidization event *T. turgidum* and *Aegilops tauschii* (DD genome, 7 chromosomes) together formed the modern hexaploid bread wheat (AABBDD genome, 21 chromosomes) 10.000 years ago (figure 1) (IWGSC., 2014, Pont and Salse., 2017). The genome contains a high proportion (>80%) of highly repetitive transposable elements (IWGSC., 2014). Even though large polyploid genomes are difficult to assemble, recently a new genome of 14.4 Gb has been sequenced and assembled based on the wheat reference genome of the Chinese Spring cultivar by the International Wheat Genome Sequencing Consortium (unpublished data).

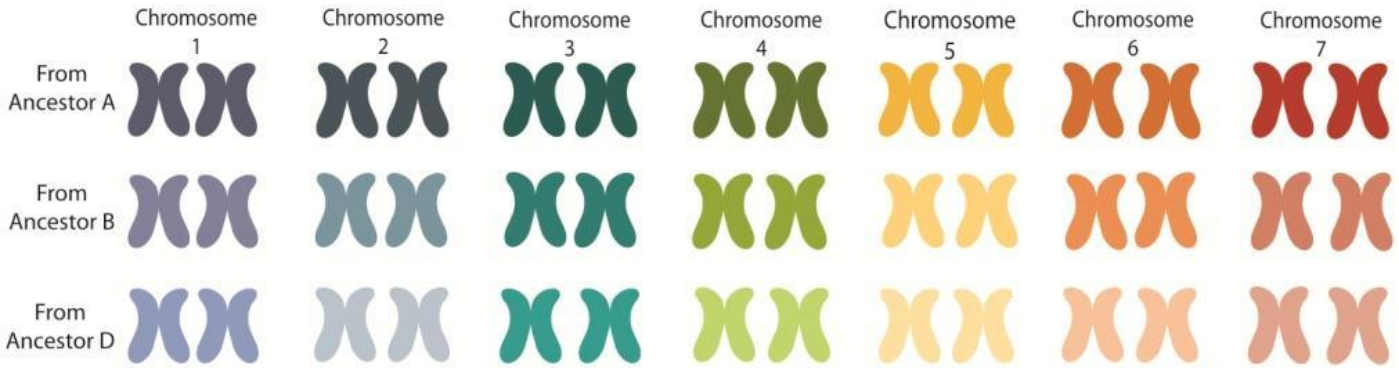


Figure 1. Representation of the wheat genome

During the course of wheat improvement many traits have been selected for, such as dwarfing and flowering time, and many disease resistance genes have been introgressed. The dwarfing trait is characterized by a decrease in stem length, resulting in shorter plants, this has been a great contributor to grain yield increase (Hedden., 2003). Dwarfing is caused by mutational alleles of the *Reduced height* (Rht) genes, which act as repressors of the plant hormone gibberellin (GA), thereby decreasing GA-responsive elongation of the stem (Hedden., 2003, Pearce et al., 2011). As wheat is grown in a wide range of different environments, flowering time is an important trait for this adaption. While winter wheat requires a long period of low temperatures to flower (vernalization), spring wheat does not and can thus be grown in spring or fall. Genes in the vernalization pathway prevent flowering in the winter, two important genes in this pathway are *VRN-1* and *VRN-2* (Yan et al., 2004). Together with vernalization, photoperiod (daylight length) response is key for heading time of wheat, which must be adjusted to the growth conditions of the crop (Yasuda and Shimoyama, 1965). For instance, heading too early could cause frost injury during winter. An example for involvement in photoperiodism is the photoreceptor gene *PHYTOCHROME C (PHYC*). Loss of function mutations in this gene showed an extreme delay in flowering, indicating the importance of this gene (Chen et al., 2014).

Among other processes, such as protein modifications, gene regulation differences are often the basis for crop improvement (Doebley et al., 2006, Swinnen et al., 2016). A review on 60 known domestication and improvement genes, revealed that most mutational changes were in regulatory genes or regions, while only 10 genes were due to protein modifications (Meyer et al., 2013). Although this is only a small selection of such genes, it illustrates the importance of regulation differences involved in plant domestication and improvement. Breeders have selected for the best crops, thereby unintentionally selected for changes in gene expression and regulatory networks. These gene regulation differences that cause phenotypic differences in crops can be caused by mutations in regulatory elements. For instance, allelic differences in trans-acting factors, affecting the expression of many downstream genes. Another possibility is allelic differences in upstream regions of genes, often promotor regions, acting on the expression of the gene or genetic region itself (cis-acting). An example of trans-acting genes regulating a trait is kernel color in maize, in which different alleles of several transcription factors of two different pathways cause different kinds of kernel colors (Ford., 2000). An example of cis-acting variation is the *VRN-1* gene in wheat.

Deletions in the promotor region of this vernalization gene are associated with spring growth as these deletions are absent in winter wheat (Yan et al., 2004).

For breeding it is difficult to fix many alleles if they are genetically unlinked, due to recombination they will segregate, and thus the chance that the offspring carries favorable alleles lowers. As an example, in a cross between two cultivars with a single additive effects locus, one quarter of the F2 population would have a favorable allele for a trait (AA). In case of two loci the ratio of obtaining homozygous favorable alleles would be $1/16^{th}$, and for three loci $1/64^{th}$. We therefore propose that selection would favor a simpler genetic control. In this simple control, variation contributing to trait differences is either found in single regulatory elements (trans-factors), or alternatively in upstream regions of many genetically linked genes (cis-factors). With these two regulations, there is a lower chance of change or loss of traits due to segregation of loci. This as linked cis-factors would most likely segregate together because they are closely located on the genome, and a single trans-factor affects all downstream genes, thus the chance of changing or losing a trait is lower.

An approach of investigating this simple regulation is by use of network analysis. With network analysis, usually the expression of many genes is monitored over different environmental conditions, genes that co-express over these conditions are clustered together. These co-expressed genes are considered to be involved in the same pathway (Langfelder and Horvath., 2008). To use network analysis for assessing regulation, we perform the clustering of co-expressed genes over a population that has genetic variation for different chromosome regions, causing the individuals in the population to have either parental allele for specific genetic regions. Genes can be co-regulated due to single regulatory factors, as the individuals carry either parental allele for a trans-acting factor that affect all downstream genes. Multiple trans-acting factors affecting a single biological process will be difficult to observe with network analysis, as due to segregation of the trans-acting factors, the downstream genes will not be co-expressed over the entire population. Alternatively, we can observe co-regulation of genes due to genetically linked cis-factors, i.e. genes located on the same chromosome that have genetic variation in each of their upstream regions affecting the expression of the gene itself. Again, the offspring carries either parental allele and due to being linked there is a lower chance of segregating.

To explore this genetic control and regulatory networks in wheat in relation to improvement, a recombinant inbred line (RIL) population was derived from a cross with the heirloom cultivar 'Red Fife' and modern cultivar 'Stettler' (registered in 2008). The semi-dwarf Stettler has higher yield values and protein concentration in comparison to Red Fife, while also flowering earlier and being resistant to several diseases due to introgression of disease resistance genes. The non-dwarf Red Fife is the oldest spring bread wheat cultivar in the west of Canada, originating around 1860.

Aim and objectives

This project aims to explore the involvement of gene regulatory variation in wheat improvement using network analysis. This project is part of an expression-QTL (e-QTL) project. Correlation in expressed genes caused by many unlinked loci would be hard to retain in a segregating population. We therefore hypothesize that for regulation of biological processes involved in plant improvement selection would favor 1) single master regulatory loci that control the expression of many downstream genes and/or 2) the fixation of alleles in upstream regions of linked genes that control their expression (cis-acting).

In this thesis we tested these hypotheses in wheat, using RNA-sequencing data from a segregating population derived from a cross between a recent and heritage spring wheat cultivar, respectively “Stettler” and “Red fife”. Our first objective is to identify the biological processes that differ between the parents using differential expression analysis. Hereafter we aim to identify single regulatory factors that control the expression of many genes and to identify linked genes that are controlled by variation in their upstream regions, using network analysis. With this analysis we identified groups of genes based on their co-expression, known as modules.

We arbitrarily defined a module as containing linked genes when more than 50% of the genes is located on a single chromosome. We expected to identify modules containing unlinked genes, these are regulated by single trans-regulatory elements. In addition, we expect to identify modules that contain linked genes. We expect that genes co-regulated due to linkage are positively and negatively correlated except if they are functionally related and are consistently upregulated by one of the parental alleles, i.e. have a bias towards one of the parents. This bias indicates that the linked genes have been

selected for and are not the result of random segregation. These results must be validated with e-QTL analysis to exclude that other factors, such as chromatin reorganizations, are causing these results and to identify the genetic basis of the regulation differences.

Approach

To test our expectations, we first trimmed our reads and mapped them to our genome. Hereafter, we quantified reads, followed by normalization. To identify the biological processes that differ between the parents we performed differential expression analysis, followed by gene ontology enrichment analysis and exploration of transcription factors and known selection genes. Selection genes are genes that have been identified to contribute to the improvement of wheat by previous research.

To identify single regulatory elements and linked genes with variation in upstream regions, we performed a network analysis. With this analysis we identified clusters of co-expressed genes (i.e. modules). To infer biological meaning to the modules we performed gene ontology enrichment analysis for genes within each module. This will give us insights in what pathways have been selected for to improve wheat. For linked genes with cis-regulatory variation we identified whether there was a bias towards one of the parents, which indicates selection for this region. This was done by identifying which genes in each module are differentially expressed between the parents, and if these genes are upregulated or downregulated in one parent in comparison to the other. Hereafter with gene ontology analysis, as well as pathway analysis we will get insights if these genetically variable and linked genes are also involved in the same biological pathway. We also identified transcription factors and known selection genes in wheat in our modules.

Methods

In this project we used RNA-sequencing data from a wheat double haploid RIL population of 154 individuals, this population was created by androgenesis on pollen of the F1 offspring. This population was formed by a cross of the modern wheat cultivar Stettler and heritage cultivar Red Fife. Each line of plants was represented by one pot containing five plants. From these five plants the second leaf at two-leaf stage (10 days old) was collected, cut about one cm from the base of the leaf, and RNA was extracted. This was repeated three times, with two weeks in between, and every time the arrangement of the pots was random to exclude environmental effects. The samples (5 leaves x 3 weeks) were pooled into one sample, thus of each individual one replicate is available. In addition, four replicates of Red Fife and four Stettler parent individuals (double haploid) are included in the data. RNA was collected at two-leaf stage because at this stage Stettler and Red Fife are in the same developmental stage.

Before library construction, mRNA molecules containing poly-A were purified. Reads were sequenced using Hiseq 2500 Illumina. Sequencing data encompasses 20-60 million paired end reads per individual, reads are 100 nucleotides long. Plant were grown in a growth room with 16h of daylight (21°C) and 8h of darkness (18°C), from May to August 2016. The relative humidity in the room was 70%. The plants were grown in '50/50 Sungro horticulture professional growing mix and turface' pots.

Quality control

Read quality was assessed with FastQC (Andrews., 2010), and low-quality sequences were trimmed with Trim Galore (--phred33 --fastqc --gzip --illumina --trim-n --paired --length 75) by dr. E. Raherison (Krueger., 2015). Reads were trimmed of adapter sequences prior to obtaining the raw data, however trim-galore with standard quality setting (20) still showed slightly lower quality in the end of the reads. As alignment of reads with standard quality setting overall showed a mapping percentage of 85%, we chose to compare different quality settings in order to see if the percentage of uniquely mapped reads increases. Overall a score between 20 and 30 is suggested to have the best trade-off between read loss and increase in read alignment with Cutadapt (Martin., 2011), this tool removes unwanted sequences from your reads and is implemented in Trim Galore. However, the exact quality cutoff score should be assessed for each individual study (Del Fabbro., 2013). We therefore compared the standard quality setting of 20 to a quality of 25 and 30, to see the effect on read alignment for one individual.

Read alignment

Reads were mapped with fast and sensitive aligning STAR (runThreadN 16 --limitBAMsortRAM 280000000000 --outSAMtype BAM SortedByCoordinate --outFilterMismatchNmax 2 --readFilesCommand zcat --outFileNamePrefix --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0) (Dobin et al., 2013) to the newly sequenced genome of the wheat reference accession Chinese Spring (IWGSC RefSeq v1.0), with the corresponding annotation file. Parameters were optimized by dr. E Raherison prior to the start of this project. To improve identification of novel splice junctions the two-pass alignment of STAR was used in which splice junctions are separately identified and quantified (Veeneman et al., 2015).

Data pre-processing

Aligned reads were pre-processed, first with Samtools (Li., et al., 2009) the aligned single end reads were filtered out and subsequently the remaining aligned paired end reads were sorted by coordinates (samtools view -h -q 255 -f 0x2 $input | samtools sort –o $output). Hereafter read duplicates, most probably originating from PCR, were filtered out with picard (java -Xmx12g -jar picard.jar MarkDuplicates INPUT=$input OUTPUT=$output METRICS_FILE=$intermediate_file). The output was again sorted by coordinates with samtools (samtools sort $input –o $output). This is necessary for usability by the read quantification program, described later.

Assessment of unmapped reads

Unmapped reads from STAR were processed with DIAMOND (Buchfink et al., 2015) and MEGAN6 (Huson et al., 2016) to understand if the lack of mapping was due to contamination or other technical reasons such as missing genome sequences. DIAMOND is a tool that uses blastx based on indexing, but is much faster. Using this tool, we performed blastx with unmapped reads against the NCBI non-redundant (nr) database (diamond blastx --threads 12 -d nr.dmnd -q inputname -o outputname.m8). Hereafter the taxonomical content of the output was be assessed with MEGAN6, using the prot_acc2tax-May2017.abin as taxonomic input, which is available at the MEGAN6 download page.

Read quantification

Transcripts were quantified with Stringtie (Pertea et al., 2015). Stringtie assembles transcripts while simultaneously estimating their expression levels, this allows better estimation of expression levels amongst transcripts than programs such as Cufflinks (Pertea et al., 2015). With the prepDE.py, available on the Stringtie manual page, raw aligned read counts were extracted from the output files of Stringtie, and arranged in a matrix (samples vs. genes). In total 110790 genes per individual were quantified.

Normalization and transformation

In R, raw quantified reads were filtered by excluding genes with less than 10 counts in 90% of the samples, resulting in 50285 expressed genes. As experimental design each line (including parents) was considered a group. Hereafter the reads were normalized for library size. For network analysis the normalized data is recommended to be transformed. The DESeq2 package in R recommends using either regularized log transformation (rlog) or variance stabilizing transformation (vst), which are both implemented in the package (Love et al., 2014). Both methods produce results on log2 scale. We therefore chose to compare these two transformations with the more conventional log2(x+1) transformation. Rlog and vst transformation use the experiment-wide trend of variance over mean in order to transform the data. These normalization methods are recommended for downstream analysis such as network analysis or principal component analysis, as these work best with gene expression data in which the variance does not depend on the mean. Eventually we chose rlog transformation as the shrinkage of low read counts is greater while also correcting more for library size in comparison to vst transformation.

Differential expression analysis parents

In addition, differential expression (D.E.) analysis between the parents was performed, with three different methods, i.e. Deseq2, EdgeR and Ballgown (q-value < 0.05). Venn diagrams were constructed with the number of identified genes and

transcripts with each method. With gene ontology (GO) analysis we identified the biological processes of differential expressed genes that matched between the DEseq2 and EdgeR method, using AgriGo v2 (fishers exact test, p value < 0.05) (Tian et al., 2017).

Gene co-expression network analysis

The 30.000 genes with highest variation over the lines were used for network analysis. We performed gene co-expression network analysis using the weighted gene co-expression network analysis (WGCNA) package in R (Langfelder and Horvath., 2008). With this method we first identified obvious outliers by sample clustering with Euclidean distance, 5 samples were excluded for further analysis (cutheight = 15). The method transforms the rlog transformed expression data into an unassigned expression Pearson's correlation matrix. This matrix is transformed into an adjacency matrix, by raising all expression values to a soft thresholding power. The identification of an appropriate soft thresholding power based on a scale-free topology fit. The adjacency matrix is in turn converted to a topological overlap matrix (TOM). An unsigned TOM measure will be used, which allows genes to cluster either both positively and negatively correlated, as transcripts can be both up or down regulated. This TOM was subsequently converted into a dissimilarity matrix. Based on this matrix, a hierarchical cluster tree is produced in which modules are identified with a dynamic tree cut procedure (Langfelder and Horvath., 2008). To identify linkage, we looked at the names of the genes within each module, as these indicate the position of the gene in the genome. Modules were visualized with the ggplot2 package in R, showing the expression pattern of the genes (scaled values) over all lines within each module as a clustered heatmap as well as a line plot.

Biological processes related to modules and differentially expressed genes parents

To identify biological function of modules and the D.E. genes of the parents we performed GO analysis using AgriGO with standard settings (Statistical test method: Fisher, multi_test adjustment method: Yekutieli (FDR under dependency), significance level < 0.05). As a background for the modules the 30.000 genes with highest variation were used, i.e. the input for the WGCNA analysis. As background for the parents we used the genes that were the input for the D.E. analysis (i.e. all genes with at least 10 counts over the parental lines). We focused on GO-terms that could possibly underlie improvement of wheat, such as dwarfing or early flowering (Peng et al., 2003). First, as our annotation file did not contain identifiers needed for functional annotation, but only gene positions, we used blastx on our entire genome against a *A. thaliana* protein database, obtained from the TAIR website (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/TAIR10_pep_20101214_updated). With a custom python scripts blastx matches were separated into gene lists with the corresponding wheat and *A. thaliana* identifier. Using R, lists of *A. thaliana* identifiers were created for each module. To perform gene ontology enrichment analysis, we used AgriGo v2 (fishers exact test, p value < 0.05) (Tian et al., 2017).

To get more insights in the biological processes we also evaluated transcription factors identified in the parents as well as the modules. This was done by matching a transcription factor database (http://arabidopsis.med.ohio-state.edu/Downloads/AtTFDB.zip) of A. thaliana to our identifiers. In addition, using literature, known selection genes were identified in our genome by using blast. Per module and in the D.E. genes we identified matches with these selection genes.

Transcriptional bias

We tested for each module if the genes within a module were biased towards in of the parents, i.e. if the genes are all upregulated or downregulated in comparison of one parent to the other. This was done by assessing if the genes within a module occur in the differentially expressed gene list of the parents. We observed if the genes were all up or downregulated towards one of the parents.

Pathway analysis

To identify secondary metabolite pathways, we used the tool PlantiSmash (Kautsar et al., 2017). This tool first predicted all genes which could be part of a secondary metabolite pathway that are clustered together on the genome. Hereafter, using R, we identified PlantiSmash genes present in each module or in the D.E. gene list. In addition, we identified metabolic pathways for several modules and the D.E. genes with BlastKoala (Kanehisa et al., 2016).

Results

The objectives of this thesis were to identify the biological differences between the parents Stettler and Red Fife, and by use of network analysis to identify either trans-acting factors and/or linked genes with variation in upstream regions affecting their expression. In order to achieve these objectives, we first had to quantify our RNA-sequencing data, that included assessing read quality, aligning to a genome, quantifying the reads and normalizing. In addition, for our data to be used for network analysis, we performed transformation of the normalized reads.

Transcript quantification

At the start of this thesis the raw read quality had already been assessed by dr. E. Raherison. Overall the read quality was high, however in order to see if we could improve the read alignment rate, which was around 84%, we tested different quality setting cutoff values with FastQC, i.e. 20 (original), 25 and 30. The difference in percentage of uniquely aligned reads was minimal between the different cutoff values, i.e. less than 0.05 % difference (see table 1). In addition, a study assessing the effect of these three quality settings (20, 25 and 30) found that the percentage of correctly mapped reads inside gene models decreases with the higher quality-cutoff trimming setting, indicating that more trimming is unfavorable (Del Fabbro et al., 2013). In addition, this study suggests that modern aligners are able to overcome potentially low quality issues, making trimming not necessary (Del Fabbro et al., 2013). Therefore, we chose to continue to work with reads trimmed with a cutoff value of 20, as these were already performed on all samples prior to the start of this project.

Table 1. Effects of different quality score cutoff values on STAR 1-PASS alignment, and effect on splices between STAR 1-PASS and STAR 2-PASS alignment.

		STAR 1-PASS			STAR 2-PASS
	Quality score cutoff	**20**	**25**	**30**	**20**
Sample 2271	Number input reads	22451643	21743139	20170557	
	Unique mapped reads	18883664	18293730	16960027	
	Unique mapped reads %	84.11%	84.14%	84.08%	
sample 10	Number input reads	19028226	18724729	18099298	19028226
	Unique mapped reads	16074932	15825191	15295973	15834963
	Unique mapped reads %	84.48%	84.51%	84.51%	83.22%
	Average mapped length	195.54	195.85	196.03	195.53
	Number of splices: Total	10398112	10280318	9973095	10800060
	Number of splices: Annotated	9734800	9625187	9338454	10788674
	Number of splices: GT/AG	10235218	10119442	9817033	10471377
	Number of splices: GC/AG	123038	132546	128720	264618
	Number of splices: AT/AC	5828	5773	5638	15613

For alignment STAR 2-pass was used, to identify novel splice junctions, useful for transcript identification in further future work of dr. E. Raherison. As can be seen in table 1, the percentage of uniquely mapped reads decreases slightly (from 84.48% to 83.22%) when using the 2-pass. However hereby the total number of splice sites increases. The total number of splices increases from 10.398.112 to 10.800.060, an increase of approximately 400.000 identified splice sites.

Overall the proportion of uniquely mapped reads was approximately 83-84%. To see if unmapped reads did not map due to contamination or due to other artifacts, such as genome incompleteness, we blasted the single end unmapped reads from 4 random samples with Diamond and visualized the output with Megan6 (see figure 2). From the results, we see that one quarter of the reads is probably bacterial contamination and small traces of fungi are present. The rest is divided in either *Fabaceae* or *Poaceae*.

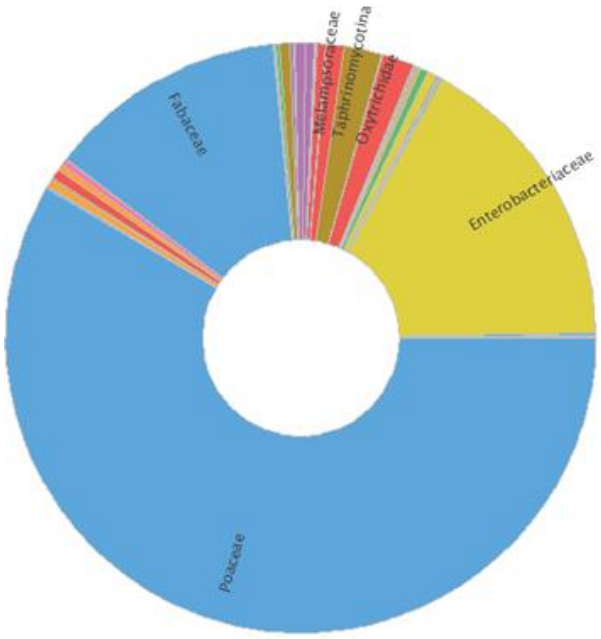


Figure 2. Distribution of identified taxonomic families in unmapped reads by MEGAN6.

After quantification with Stringtie, generated raw reads count tables were normalized with R package DESeq2. First, we performed our analysis with 18 samples, including both parents, in order to test our pipeline. As for network analysis either regularized log (rlog) or variance stabilizing (vst) transformation was recommended by DESeq2 instead of the more conventional log2 transformation, we decided to compare these methods. We plotted two random samples against each other, using the different transformation methods, first the common log2 transformation, against rlog and vst (figure 3). From this plot we can see the rlog and vst transformation correct more strictly than the log2 for low count values. Comparison data transformation methods (over 18 samples), we will use rlog as it corrects most for low read as well as high read counts.

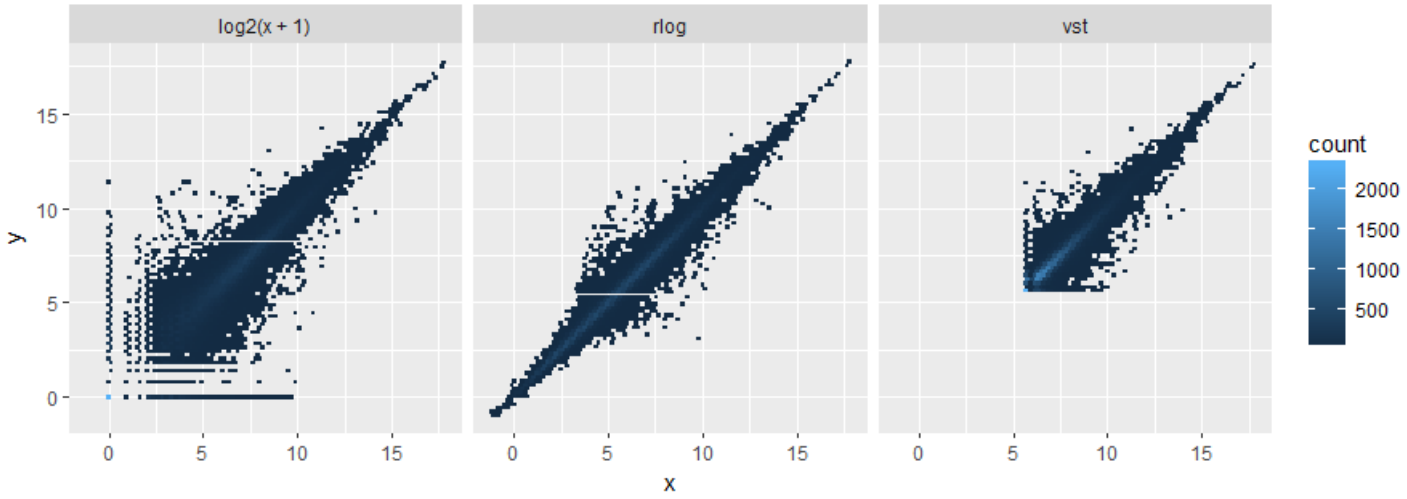


Figure 3. Two random samples plotted against each other, with three different transformation methods, i.e. log2, rlog and vst. The expression values are for all three methods on log2 scale.

Hereafter, all 162 samples were transformed, however the dispersion plot, a step used within both rlog and vst transformation, looked unusual (see figure 4A and 4B). In this step the estimate of each gene is shrunk towards a fitted value (dispersions dependence on the mean), shown with a red line. Using only 18 samples, this plot looks as expected, however using all 162 samples the red line shows a strong distortion. Removing samples identified as outliers by further steps (91, 155, 184 and 186) did not increase the quality of this plot. Possibly the high number of samples gave problems during the fitting of estimates. We still continued with this data as no clear reason was identified.

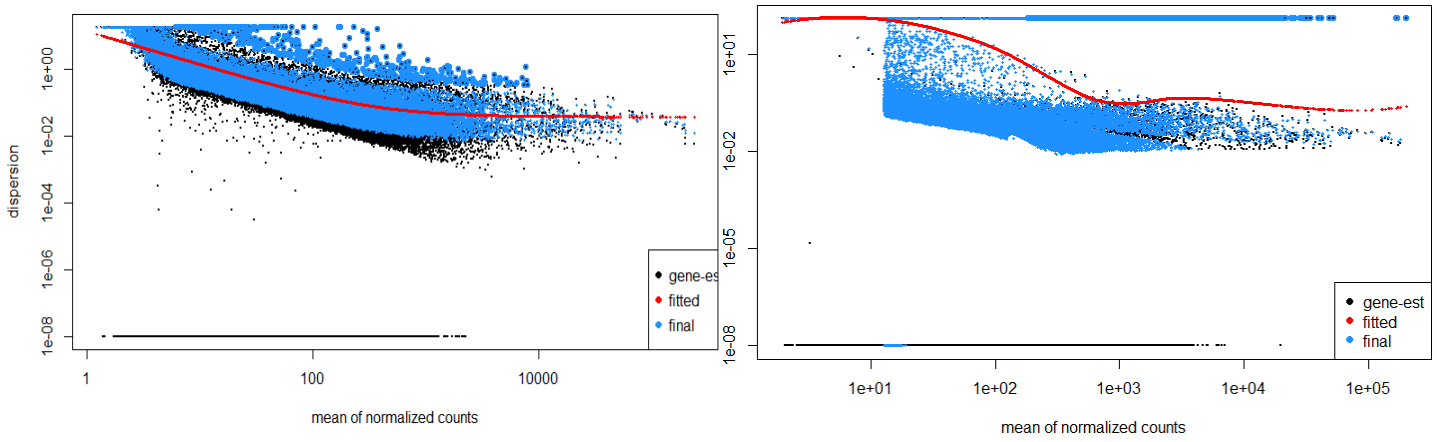
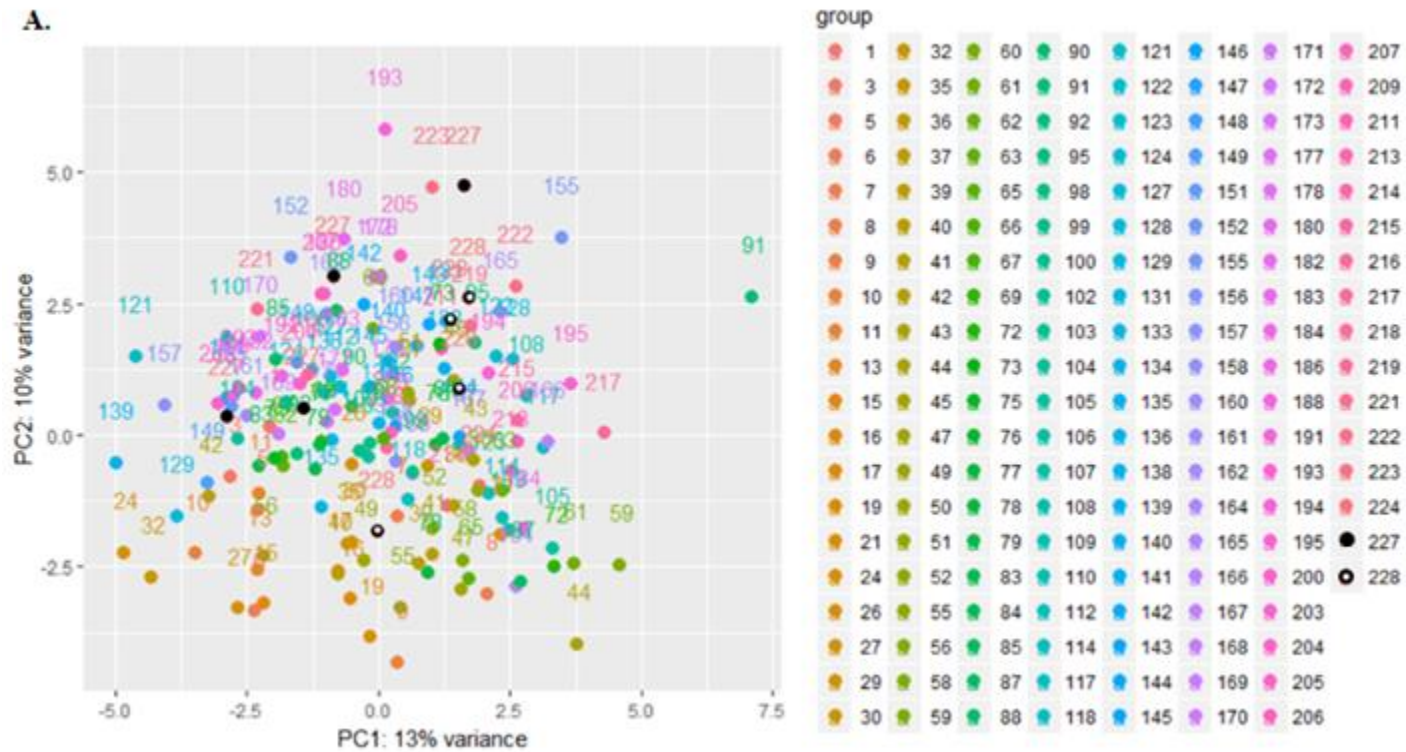


Figure 4 A-B. Dispersion plot. The black dots indicate gene-wise estimates. The red line indicates the fitted estimates. The blue dots indicate the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are not shrunk towards the fitted value due to being flagged as outliers. **A. Including 18 samples. B Including all 162 samples.**

For exploration of our data a principal component (PC) plot was performed on all genes in the rlog transformed samples (figure 5A-C). In total 162 PCs were calculated, as we use our lines as grouping factor. We expected the parents to be on either side of the plot, with all offspring lines in between. For PC1 and PC2, as expected, the lines are not clustered into groups, but distributed over the whole plot (figure 5A). The parents (227 and 228) are not strongly clustered together, but still quite distributed between all other samples. The variance of the first two principal components is low, i.e. 13% for PC1 and 10% for PC2. However, when looking at PC3 and PC4, we see a strong clustering of the parents, to either side of the plot, with as expected all lines in between (figure 5B). The variance of PC3 and PC4 is respectively 5% and 4%. This random distribution seen in figure 5A, is possibly due to an external factor, such as technical noise due to pre-processing or an environmental effect. The variance of the first 8 principal components is visualized in figure 5C, each PC has very low variance percentages.



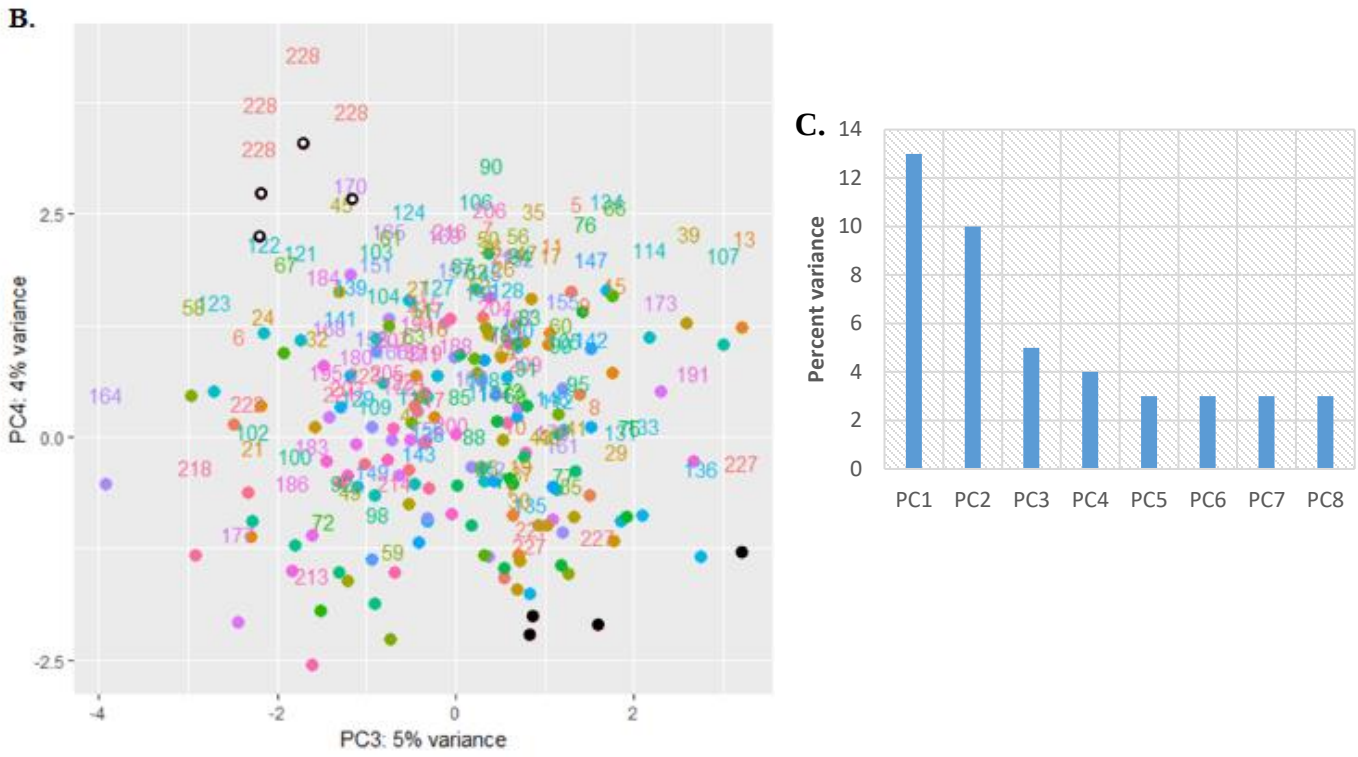


Figure 5A-C. Principal component plot of all 154 lines and the parents. 227 indicates the heritage cultivar parent "Red fife" (black dot), 228 indicates the modern cultivar parent "Stettler" (black dot with white center). **A. PC1 vs. PC2. B. PC3 vs. PC4. C. Percentage of variance explained by the first 8 principal components.**

Differential expression analysis between parents

To identify which biological processes differ between the parents, the heritage cultivar Red Fife and modern cultivar Stettler, we performed a differential expression analysis followed up by GO analysis. We first performed differential expression analysis with three different methods; DESeq2, EdgeR and Ballgown (with q-val < 0.05) (figure 6A). DESeq2 and EdgeR show a big overlap of differentially expressed genes, i.e. 6039 (figure 6B). The overlap with Ballgown however is smaller, here only 2568 genes are identified. We chose for further analysis to work with genes identified with both DESeq2 and EdgeR, this to increase robustness and not rely on one method. For visualization of significantly differentially expressed genes, a MA-plot and volcano plot was made with D.E. genes of DESeq2 (figure 6C and 6D).

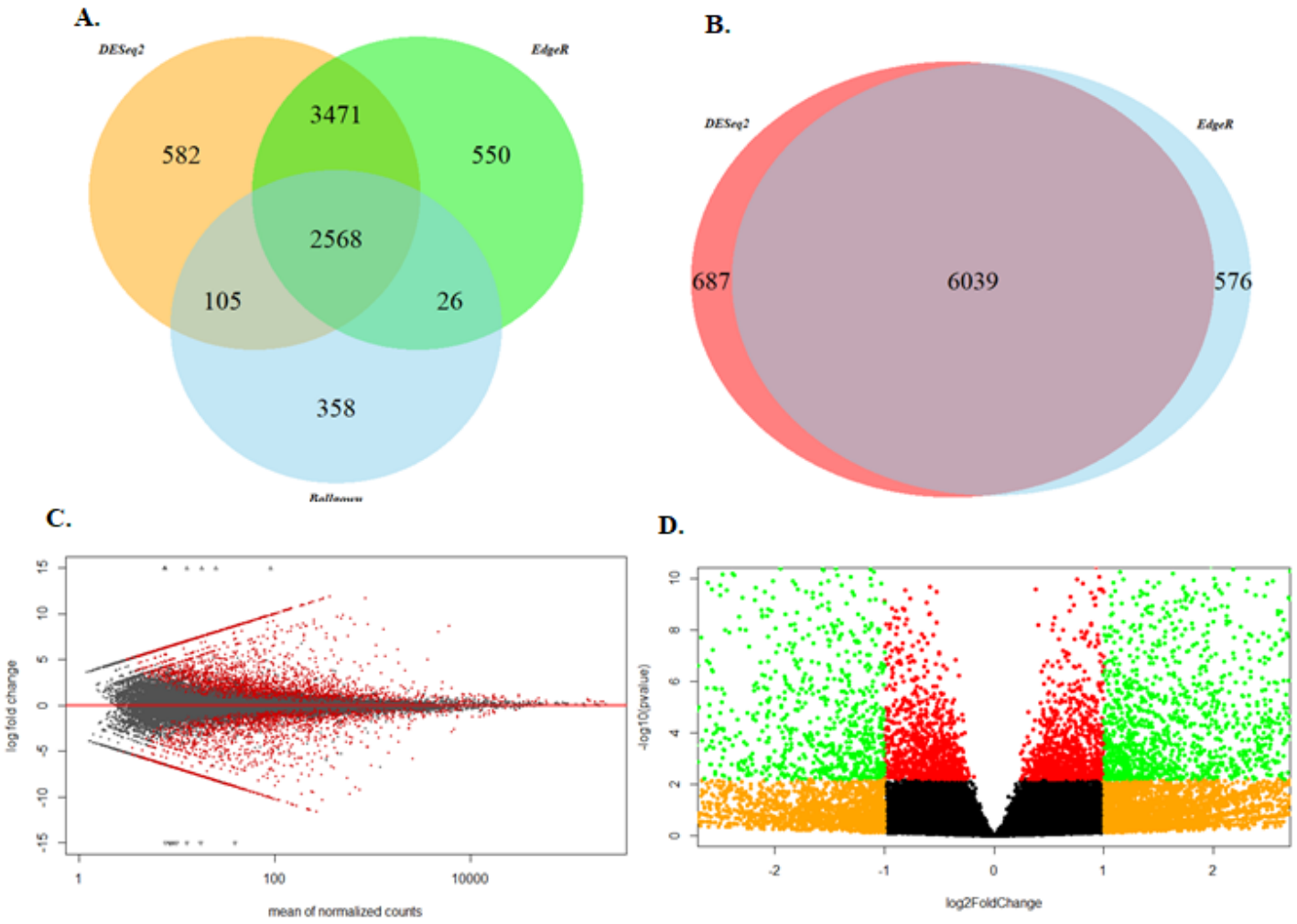


Figure 6 A-D. Differentially expressed genes in DESeq visualized between parents. A. Comparison of identified differentially expressed genes between DESeq2, EdgeR and Ballgown. B. Comparison of identified differentially expressed genes between Deseq2 and EdgeR. C. MA-plot. This plot shows the mean average of genes against log fold change. Values in red show genes which are significantly differentially expressed, with an adjusted p-value smaller than 0.05. **D. volcano plot**. This plot shows the log2foldchange of genes between the parents. Values in orange show genes with a log2 fold change bigger than 1, Values in red show genes with an adjusted p-value smaller than 0.05. Values in green are the intersect of red and orange genes.

To identify biological processes differing between the parents, we performed GO enrichment analysis, with genes differentially expressed in both EdgeR and DESeq2. As there was no wheat annotation yet for our genome, we used *A. thaliana* identifiers, obtained through blastx, which represented 85.7% of our wheat genes (5176 Arabidopsis genes in 6039 wheat genes). We first used all differentially expressed genes to identify significant GO terms (p-val <0.05 and FDR<0.05), this gave as a result in total 21 terms mainly related to defense response and response to external factors such as stress or biotic stimuli (see table 2).

Table 2. GO-analysis results with all differentially expressed genes between parents.

GO term	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006952	defense response	287	732	7.50E-14	7.80E-10
GO:0006950	response to stress	604	1985	5.30E-10	2.70E-06
GO:0050896	response to stimulus	954	3408	4.50E-09	1.50E-05
GO:0006468	protein phosphorylation	208	559	8.30E-09	2.10E-05
GO:0043207	response to external biotic stimulus	208	583	1.10E-07	0.00016
GO:0009607	response to biotic stimulus	210	589	1.00E-07	0.00016
GO:0051707	response to other organism	208	581	9.10E-08	0.00016
GO:0016310	phosphorylation	259	765	1.20E-07	0.00016
GO:0098542	defense response to other organism	157	428	1.10E-06	0.0013
GO:0044550	secondary metabolite biosynthetic process	69	144	1.70E-06	0.0017
GO:0044699	single-organism process	1379	5316	2.50E-06	0.0023
GO:0019748	secondary metabolic process	97	235	3.00E-06	0.0025
GO:0051704	multi-organism process	260	808	3.20E-06	0.0025
GO:0009605	response to external stimulus	254	792	5.10E-06	0.0035
GO:0009617	response to bacterium	113	291	4.80E-06	0.0035
GO:0042742	defense response to bacterium	94	243	3.30E-05	0.021
GO:0009699	phenylpropanoid biosynthetic process	38	71	5.90E-05	0.031
GO:0006796	phosphate-containing compound metabolic process	321	1080	5.30E-05	0.031
GO:0006793	phosphorus metabolic process	323	1090	6.10E-05	0.031
GO:0044710	single-organism metabolic process	657	2410	5.60E-05	0.031
GO:0044711	single-organism biosynthetic process	313	1057	8.30E-05	0.041

To assess if we could define more GO-terms we decided to separate the gene set in upregulated and downregulated genes for Stettler in comparison to Red Fife. Performing GO analysis with only the upregulated genes resulted in a larger number of significant GO-terms, i.e. 27 (see Appendix table 4). The majority of the terms include response to external effects, such as stress or biotic stimuli. In addition, there are significant GO-terms related to the immune response and salicylic acid. Downregulated genes showed only 5 significant GO terms, again related to defense response and external factors (see Appendix table 5).

Hereafter we looked at known transcription factors in *A. thaliana* and identified 229 of these transcription factors in our differentially expressed gene list (see appendix table 6). Assessing the biological processes of these transcription factors resulted in many relevant terms related to the improvement of wheat. For instance, several terms related to defense response and abiotic stress, e.g. involvement of the stress hormone abscisic acid. Some transcription factors involved in defense response include *WRKY DNA BINDING PROTEIN 40 (WRKY-40)* and *70 (WRKY-70)*. The biological process flower development was also observed several times, e.g. in *APETALA1 (AP1), EARLY FlOWERING MYB PROTEIN (EFM), NGATHA1 (NGA1)* and *AGAMOUS-LIKE 14 (AGL14).* Several *A. thaliana* transcription factor homologs related to leaf morphogenesis were identified, such as *AUXIN RESPONSE FACTOR 3 (ARF3)*. Also, transcription factors related to gibberellin (GA) pathways were identified, such as *BETA HLH PROTEIN 93 (NFL)* and *REPRESSOR OF GA (RGA1)*, a member of the DELLA regulatory family, which could indicate involvement in dwarfing. For the described transcription factors, we assessed if the encoding genes are upregulated in the modern Stettler or heritage Red Fife (Appendix table 7). *WRKY-40, RGA1, EFM, ARF3, NGA1* and *WRKY-70* were all upregulated in Stettler in comparison to Red Fife, while *AP1, AGL14* and *NFL* were upregulated in Red Fife.

To identify possibly secondary metabolite pathways, we used the tool Plantismash, and compared how many of the genes from predicted pathways are present in our DE gene list. Plantismash predicted in total 236 clusters, including in total 2987

genes. From our differentially expressed gene list, we identified 189 Plantismash genes originating from 111 clusters, however it is important to look if full pathways were identified or only single genes, the latter would be less meaningful. No full secondary metabolite pathways were identified, mostly 1or 2 genes of a full pathway were found, usually less than 20% of a cluster. The highest observed percentage of genes identified in a cluster was 44.4%, i.e. 5 out of 11.

We also used BlastKoala to assess possible pathways in our gene list. A lot of different processes were identified (see figure 7). For instance, processes related energy metabolism, such as photosynthesis (7 Kegg identifiers). One Kegg identifier was identified that is involved in diterpenoid (gibberellin) biosynthesis, categorized under metabolism of terpenoids and polyketides. In total 16 KEGG identifiers belonging to Plant hormone signal transduction were identified, in which 2 proteins involved in gibberellin transduction and 3 in abscisic acid downstream signal transduction were found (figure 8). We assessed again the log2fold change between the parents for the genes shown in this figure (appendix table 8). The DELLA repressor is upregulated in Stettler, while the repressor of DELLA, *GID2*, is upregulated in Red Fife. DELLA represses further downstream signaling of gibberellin, which could be important for the reduced stem growth in Stettler. For the ABA downstream signaling, *PP2c* and *SnRK2* are upregulated in Stettler, while *ABF* is upregulated in Red Fife.

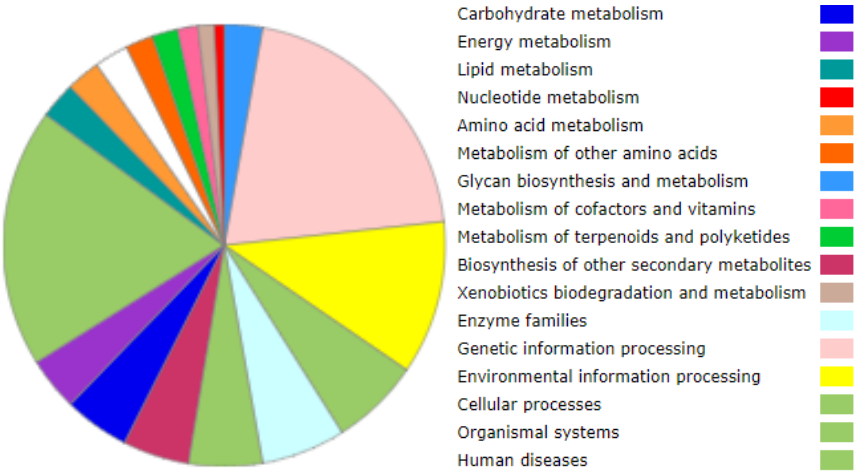


Figure 7. Functional categories identified with BlastKoala

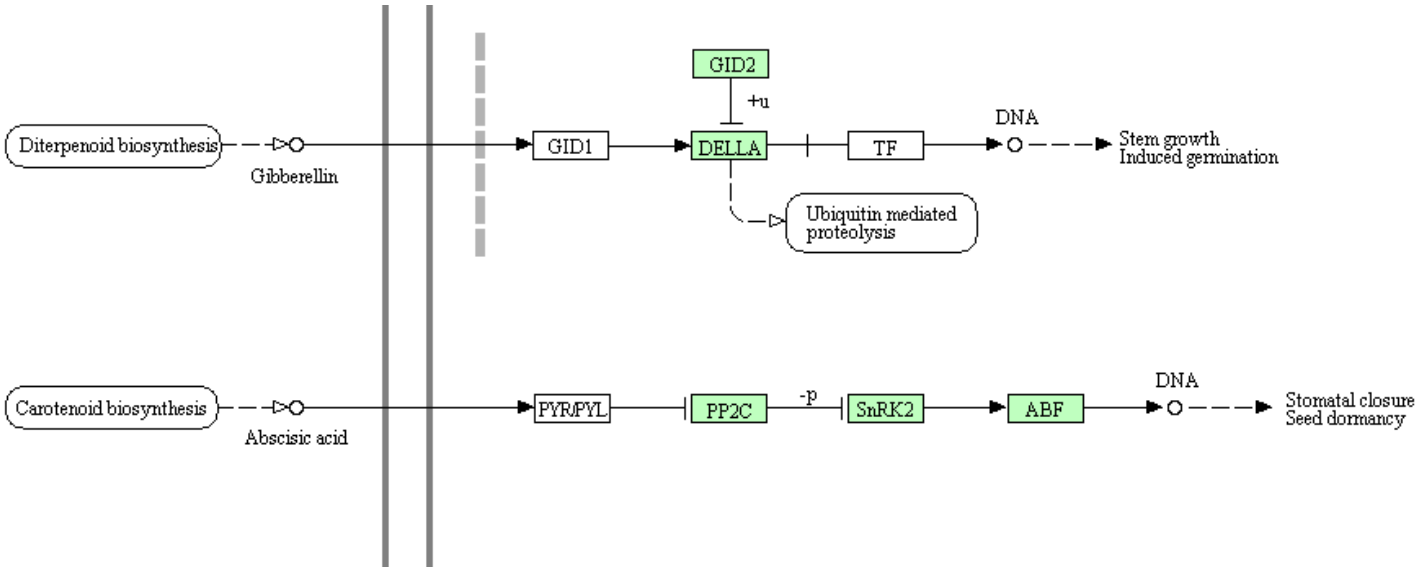


Figure 8. Two examples of identified components by BlastKoala in Plant hormone signal transduction pathways. With gibberellin and abscisic acid downstream signaling.

We also looked if already known selection genes were present in the list of differentially expressed genes. For this we first identified the matching genes in our genome. Two of these genes were identified in our gene list, i.e. *RHT-B1 (I7HUS4)* and *PPD-D1c (A7J5U2). RHT-B1* encodes a GA-insensitive protein, involved in dwarfing which is a common selection

trait in wheat. This gene is identified also as the DELLA repressor by BlastKoala and as *RGA1* by blastx on *A. thaliana*. The PPD-D1c protein is overall associated with later flowering, and is non-responsive to photoperiod. This gene is upregulated in the heritage cultivar Red Fife, this is in line with the fact that Red Fife flowers later than Stettler.

Gene co-expression network analysis

To explore the genetic basis behind transcriptional variation between Red Fife and Stettler we performed a network analysis. This was followed by a GO-analysis to assign biological processes to modules. For the network analysis we used 30.000 genes with highest variation over samples after rlog transformation. First, we detected outlying samples by sample clustering (see Appendix figure 15). Samples higher than the cutoff height of 15 were discarded for further analysis, i.e. line 188, 91,155, 193 and one of the parental samples 227. After gene clustering and assigning modules 94 modules were generated (figure 10). We decreased the number of modules by cutoff of clustered eigengenes (Appendix figure 16), modules below this cutoff were merged, resulting in a final module number of 73, including an unassigned grey module (see figure 9). The average size of our modules was 416.7 genes, with our smallest module including 30 genes and the largest module including 5813 genes.

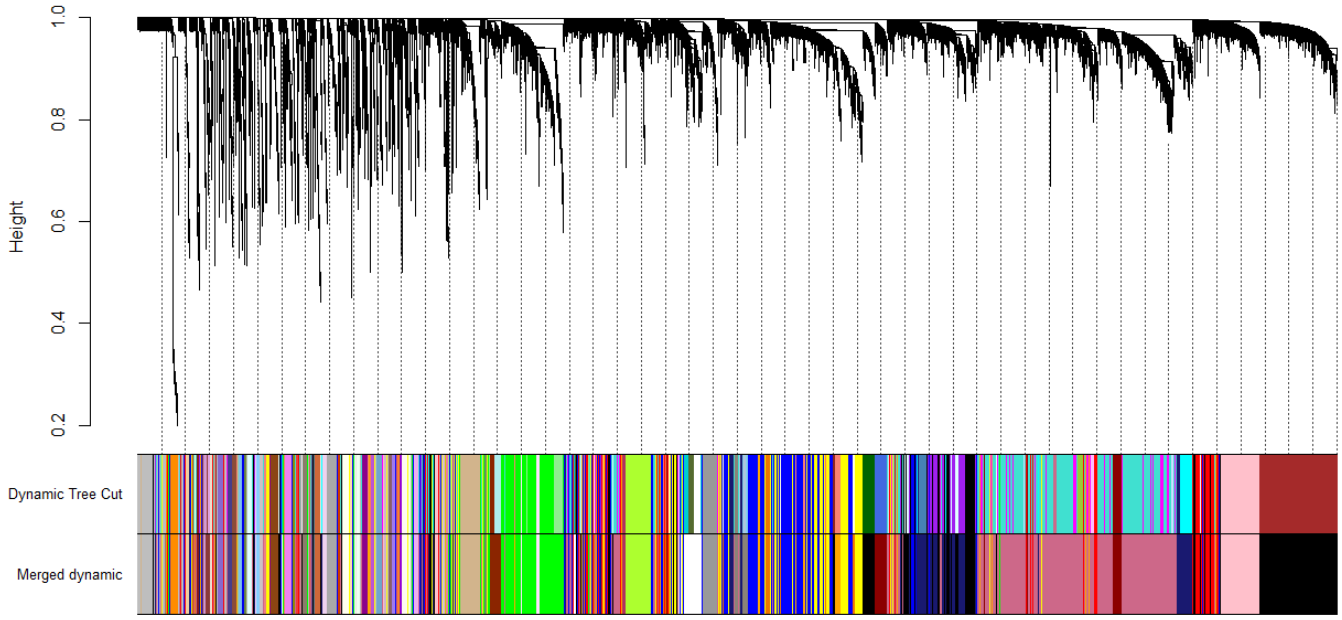


Figure 9. Gene cluster dendrogram created by WCGNA, visualizing modules before and after merging of eigengenes.

Our objectives were to identify modules in which genes are regulated by single trans-acting factor, and modules in which linked genes occur that each are regulated cis-regulatory variation. Genes co-regulated by a trans-acting factor will likely be positively co-regulated and not linked. These genes will be clustered into a module due to co-regulation. In addition, if linked genes vary due to cis-regulatory variation one expects them to cluster in a module. Genes co-regulated due to linkage are expected to be positively and negatively correlated unless they are functionally related (i.e. involved in the same pathway) and consistently upregulated by one of the parent's alleles. To determine if modules are composed of linked or unlinked genes, an arbitrary threshold was chosen assigning a module as linked if more than 50% of the genes is located on a single chromosome, or unlinked if less that than 50% of the genes is located on a single chromosome. We indeed identified indeed in total 26 modules with only little genetic linkage, while 46 modules showed linkage of genes. A summary of module information is shown in table 3, more extensive information can be found in appendix table 9.

Table 3. Overview of module information.

Total number of modules (excluding unassigned grey module)	72
Average size	416.7
Number modules without linkage	26
Average percentage of modules with no linkage (<50%) with only positive correlation	66.7 %

Average percentage of modules with linkage (>50%) with positive and negative correlation	97.8 %
Percentage of modules with unlinked genes (<50%) with significant GO-terms	65 %
Percentage of modules with linked genes (>50%) with significant GO-terms	22.2 %
Total number of D.E. genes of parents in modules	4587
Average number of transcription factors in modules	23
Average representation of *A. thaliana* identifiers	88.1 %

We visualized the expression of genes in all modules, by plotting their scaled (subtracting means and dividing by standard deviation per gene) expression patterns over all lines in a line graph as well as a clustered heatmap. In many modules we observe both negative and positive correlation (for example see figure 10A). In 19 modules there is clear positive correlation, with only few genes showing negative correlation (for example figure 10B). This can also clearly be observed in the clustered heatmaps made for each module, see figure 10C and D. We observed that most modules without linked genes had a positive correlation expression pattern (66.7%) (figure 10B), while almost all modules with linkage have both positive and negative correlation (97.8%) (10A). In modules showing positive and negative correlation, the parents are in most cases clearly clustered to either side of the pattern, i.e. each parent has either lower or higher expression. In the unlinked modules we see less strong clustering of the parents, they are often grouped together but not clearly separated in having all higher or lower expression values. Hereafter the number of differentially expressed genes between the parents found back in modules was assessed. In total 4587 genes were identified in modules. This means that 1452 genes which are differentially expressed between the parents were not found back in modules.

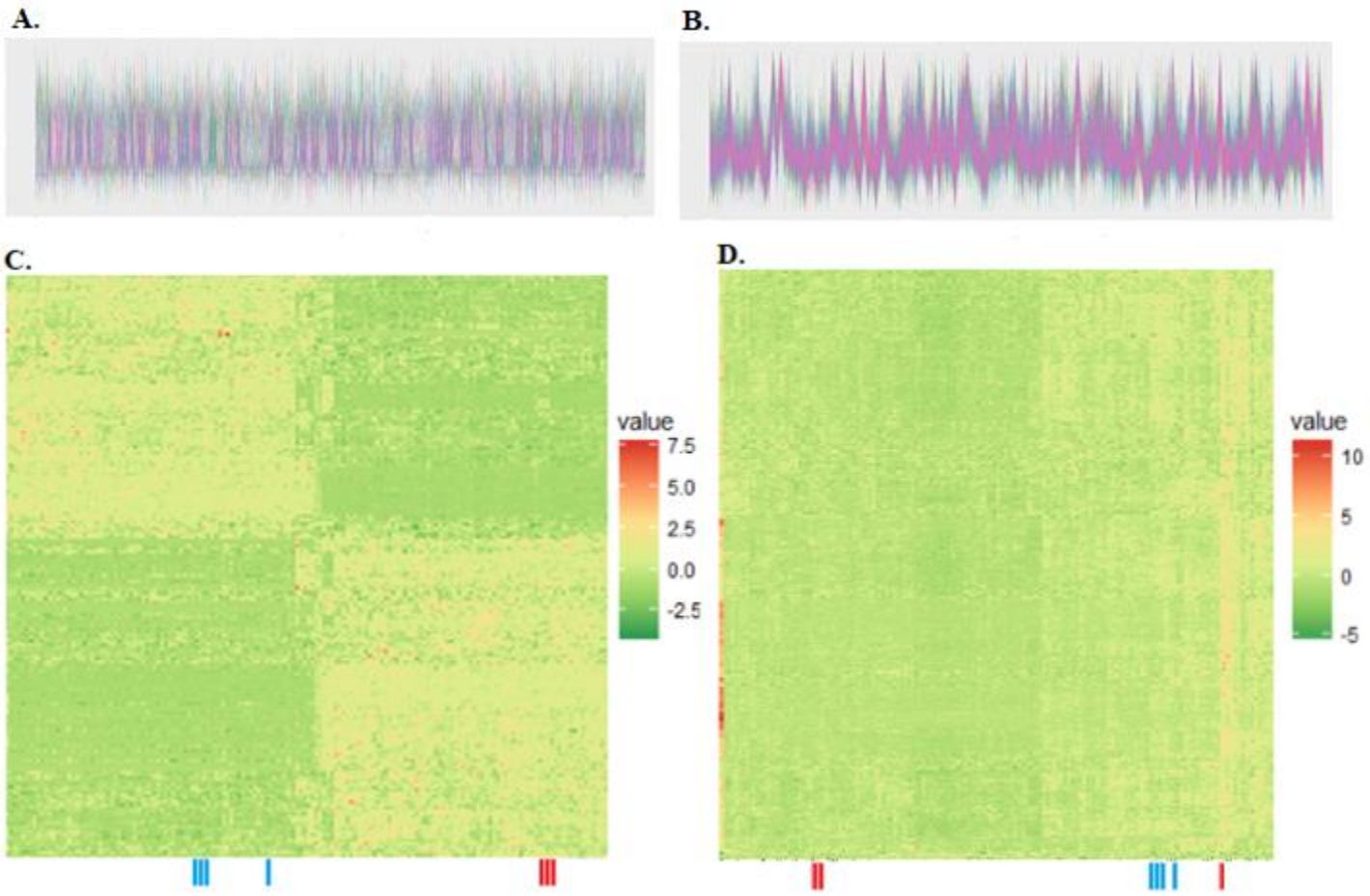


Figure 10. A-D. Correlation patterns and clustered heatmaps. For all figures the x axis indicates the different samples and y axis either scaled expression values (A+B) or gene names (C+D). Blue and red indicate the parents, respectively modern Stettler and heritage Red Fife **A. module saddlebrown, both positive and negative correlation b. Module tan, clear positive correlation over the module c. Clustered heatmap saddlebrown. D. Clustered heatmap tan.**

A boxplot was made to visualize more clearly the positive and negative correlation observed in modules with linked genes and the positive correlation in modules without linked (figure 11). In this boxplot the average correlation values of the genes were calculated and separated in modules with linked genes and without linked genes. Modules with linked genes have a clear lower average correlation, most probably due to having both negative and positive correlation of genes, while modules without linked genes have a higher average correlation due to having mainly positive correlation between the genes.

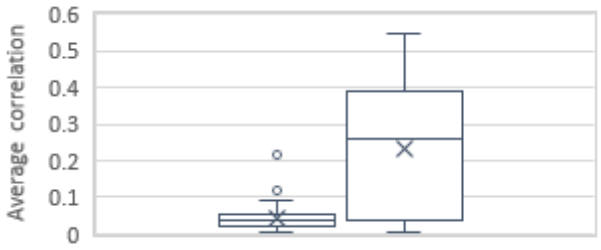


Figure 11. Average correlation of genes in modules with linked genes (left) and modules without linked genes (right).

Modules without linked genes

In this section we will describe in more depth the results for modules without linked genes, which are potentially regulated by single trans-acting factors. We assessed the biological processes related to modules to see which processes are related to the improvement of wheat. A small number of interesting modules will be described in more detail, chosen based on being potentially involved in the improvement of wheat. In total 65% of the modules without linked genes had enriched GO-terms. Several of these modules have enriched GO-terms related to defense response and response to external factors, such as abiotic or biotic stress. Other enriched terms potentially involved in wheat improvement include photosynthesis and flower development.

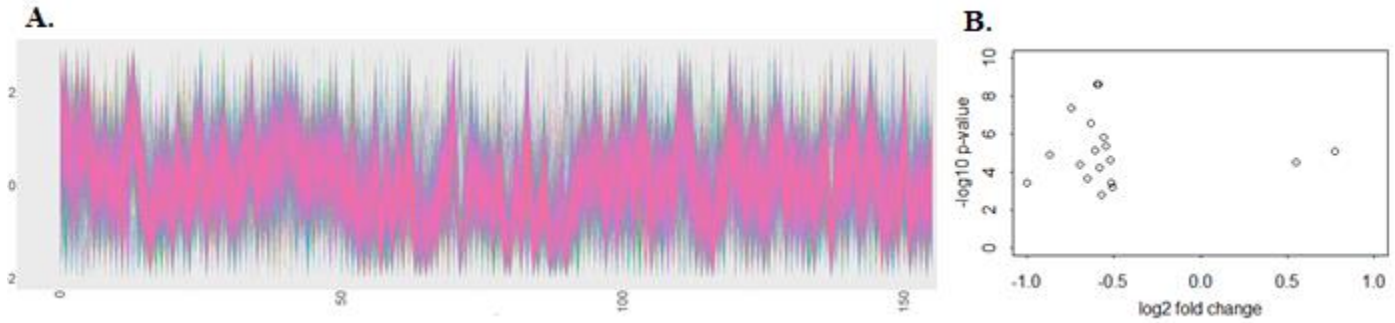


Figure 12. A-B. A. Expression pattern of pink module over all samples. B. Genes from pink modules that significantly differ between the parents plotted (Red fife vs Stettler), with p-value < 0.05. Shows bias towards ancestral cultivar Red Fife. Visualized with genes from the pink module which are found back in the differentially expressed gene list of parents.

The pink module (figure 12A), with a size of 1099 genes, encompassed several GO terms related to flowering, epigenetics and differential splicing (see appendix figure 10). In addition to this, a known selection gene encoding wheat phytochrome and flowering time protein 1 was found in this module. This protein is involved in flowering time, inducing flowering under optimal light conditions, while, mutants cause late flowering. In total 45 transcription factors were identified in this module, many of these are involved in processes such as flower development and epigenetics, such as chromatin silencing. A small selection of these transcription factors can be found in Appendix table 11. Most of the pink module genes identified in the D.E. gene list shows a bias towards the ancestral parent Red Fife, as these genes are significantly downregulated in Stettler (figure 12B). Stettler flowers earlier than Red Fife, thus possibly Red Fife carries also some early flowering alleles or possibly these genes induce flowering time. However as only 18 genes out of the 1099 genes present in the module are visualized it might not be a good representation.

The green module (figure 13A) is enriched in GO-terms related to defense and (biotic) stress response, also a known selection gene (WRKY) involved in the resistance against the fungus fusarium head blight was identified in this module. In total 58 transcription factors were identified, most of these seem to be related to regulation of any external factor. We observe for instance transcription factors with biological processes related to response to chitin, fungus or bacterium, but also to salinity and heat response (Appendix table 12). In total 564 out of 1759 genes in this module were found back in the list of differentially expressed genes in the parents and show a strong bias towards the modern Stettler parent (figure 13B), indicating selection.

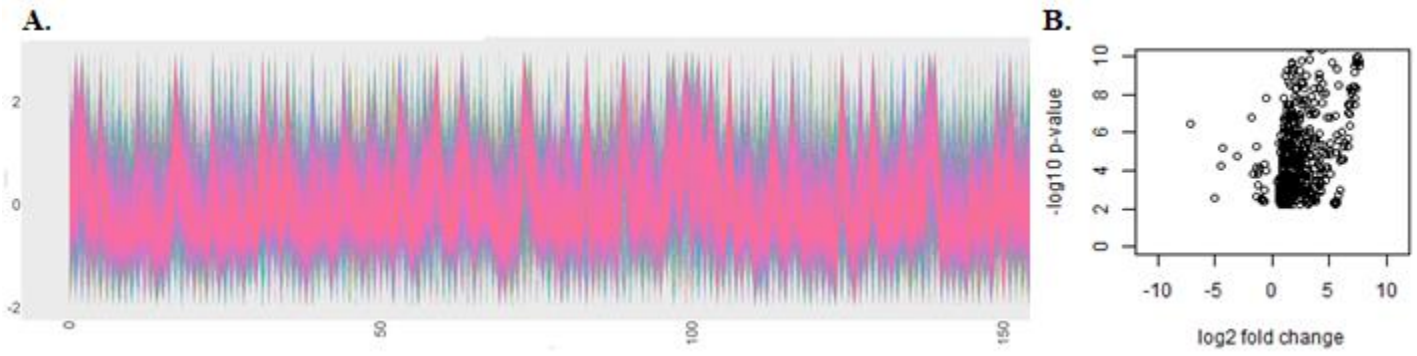


Figure 13. A-B. A. Expression pattern of green module over all samples. B. Genes from green modules that significantly differ between the parents plotted (Red fife vs Stettler), with p-value < 0.05. Shows bias towards modern cultivar Stettler. Visualized with genes from the green module which are found back in the differentially expressed gene list of parents.

Another module with interesting GO-terms is the darkorange module, including terms related to photosynthesis. Two transcription factors were identified in this module, i.e. a telomere binding protein AT1G49950 and basic-leucine zipper (bZIP) transcription factor family protein AT5G11260. The latter is involved in light-regulated transcriptional activation of G-box-containing promoters. No known selection genes have been identified in this module.

To see if modules without no linked genes include pathways with multiple KEGG identifiers we assessed four random modules, i.e. orange, blue, green and maroon. In the module orange there were multiple KEGG identifiers found within one pathway. In total 10 identifiers were found related to photosynthesis (energy metabolism), mostly involved in components of photosystem one and two. This module is also enriched GO-terms related to photosynthesis. Blue is enriched in protein localization and vesicle transport. A lot of KEGG terms related to metabolism were identified, such as energy metabolism, e.g. 8 for photosynthesis, and carbohydrate metabolism. Green is enriched in a lot GO of terms related to response to external stimuli, defense response and (a)biotic stress. Interesting pathways with multiple identified KEGG identifiers include biosynthesis of antibiotics, and signal transduction in MAPK signaling pathway. Maroon includes no significant GO-terms. The only pathway in which two hits were identified was related to plant-pathogen interaction. Other pathways were only represented by one KEGG identifier in Maroon.

Modules with linked genes

For modules with linked genes, we expect that genes are not selected for during wheat improvement, unless there is a bias towards one of the parents and the genes are involved in the same biological process. Overall, for modules with linkage 22.5% of the modules included significant GO-terms. These GO-terms were mainly involvement in phosphorylation and defense response. None of these significantly enriched modules had a transcriptional bias towards one of the parents. This is in line with a previous result, in which the parents are located on either side of heatmaps of modules with linked genes (figure 10D).

One module with linkage showing a bias towards one of the parents is module blue2 (figure 14A-C), a module with 92.5% of the genes located on a single chromosome. This is the only module, together with skyblue1, showing a bias towards one of the parents. The co-expression pattern of the genes is both positive and negative suggesting cis-factors, thus having either allele of the parents. Genes in blue2 have a transcriptional bias towards the modern cultivar Stettler, this could indicate selection (figure 14C). However, no significant GO terms were identified for this module. The transcription factors identified in this module include different biological processes, such as cell wall organization, response to stimulus, and flower development (Appendix table 13.).

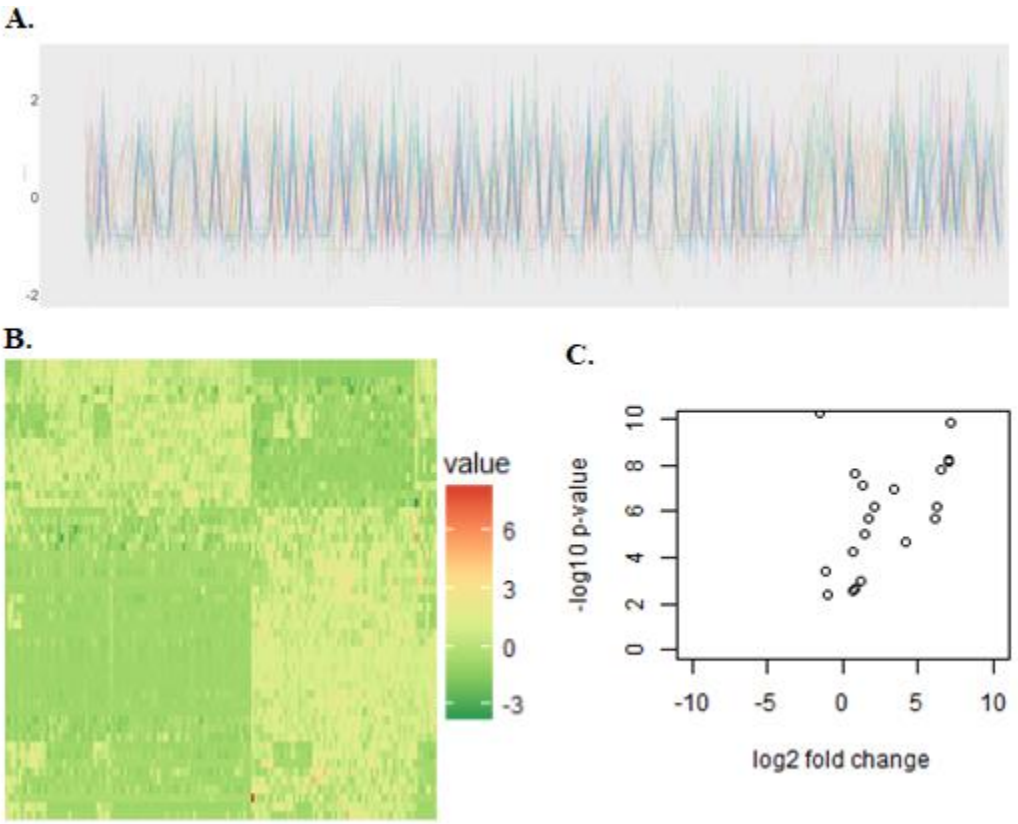


Figure 14. A-C. Information blue2 module. A. Expression pattern of blue2 module over all samples. With the sample on the x-axis and scaled expression values on the y-axis. **B. Clustered heatmap of blue2.** With the samples on the x-axis and genes on the y-axis. **C. Genes from blue2 module that significantly differ between the parents plotted (Red fife vs Stettler), with p-value < 0.05.** Shows bias towards modern cultivar parent blue2. Visualized with genes from the blue2 module which are found back in the differentially expressed gene list of parents

To get more insights if modules with linked genes are also functionally linked we looked at pathways. We used the tool PlantiSmash to identify clustered genes of secondary metabolite pathways, and see if these occur fully in our modules. With this tool in total 236 clusters were identified over the whole genome, including 2987 genes. However, no full secondary metabolite pathways predicted by PlantiSmash were identified in modules. Co-expression analysis within PlantiSmash itself did not give any result. The highest percentage of genes identified in a secondary metabolite pathway was 38% in the coral2 module, related to a saccharide biosynthesis. This module does not have any significant GO terms. Still two known selection genes were identified, encoding vernalization protein VRN-3 and Flowering time locus T-like protein 3. With BlastKoala we identified several KEGG identifiers for coral2, we observed that all components belong to different pathways, thus not identifying a full potential pathway.

We used BlastKoala on several other modules that include linked genes, i.e. blue2, skyblue1, honeydew1 and darkslateblue. Skyblue1 and Blue2 are modules without significant GO-terms, but both show a bias towards one of the parents (see figure 14C for blue2). Honeydew1 is not enriched in a GO-term but does include one known selection gene (involved in Fusarium resistance). Darkslateblue shows significant GO terms related to defense response. In all blue2, skyblue1 and darkslateblue several KEGG identifiers were identified, however again all from different pathways, representing only one identifier per pathway. In honeydew1 up to three components per pathway were identified. For instance, two components related to effector triggered immunity, which is involved in plant-pathogen interaction. Still, no full pathways are identified, indicating that different components of pathways are likely to be distributed over the genome and not genetically clustered together. Overall, we did not find strong evidence linked genes with cis-regulatory variation have contributed to the improvement of wheat, with possibly few exceptions as still two modules showed a bias in the parents.

Discussion

This thesis aimed to explore the involvement of gene regulatory variation in relation to biological processes involved wheat improvement, such as disease resistance, dwarfing, flowering time or abiotic tolerance. We performed differential expression analysis between the parents to get insights in the biological processes that differ significantly. Hereafter performed network analysis to identify single regulatory factors that control the expression of many genes and linked genes that are controlled by variation in their upstream regions.

Transcript quantification

In order to see if we could increase the percentage of uniquely mapped reads, we compared if increasing the quality cutoff parameter to 30 during trimming resulted in more uniquely mapped reads (after alignment) than with the original used quality cutoff value of 20, used prior to the start of this project. We did not see a large increase in uniquely mapped reads. This is probably because the quality of the reads was already very high when obtained (raw), thus increasing this cutoff will not resulting in better alignment.

For unmapped reads we assessed their taxonomy, in order to get insights in why they did not map. This revealed that part of it is due to bacterial and fungal contamination. But most of the unmapped reads were categorized as *Poaceae*, of which wheat is a member, or closely related *Fabaceae*. The reason for this is most probably that our genome is not fully complete, thus missing parts cause the lack of mapping.

Quantified read counts were transformed using R package DESeq2, this package suggests using either vst or rlog transformation for network analysis (Love et al., 2014), therefore we tested both methods. In comparison to the more conventional log2 transformation, rlog and vst transform the count data to a log2 scale, but reduce the variance of genes with low counts, as these are more affected by counting noise, and could thus be less biologically meaningful. The vst calculates a variance stabilizing transformation from the fitted dispersion mean and then transforms the expression count data, normalizing by the size factors, this results in a matrix of values which have approximately a constant variance along the range of mean values. The rlog is less sensitive to differences in library sizes, and thus more robust when library sizes differ (Love et al., 2014). We chose to work further with rlog transformation as it shows to corrects for low read counts while also being more robust for differences in library size.

One issue was the unexpected dispersion plot when performing our pipeline with all samples. Prior to transforming with either vst or rlog the dispersions for the gene expression matrix are calculated. The result of this, the fitted dispersions, is an input parameter for both transformation methods. Excluding outlying samples (after sample clustering) did not change the result. We could not identify a clear reason why this plot behaves as it does. Possibly the large number of samples (162), caused problems during the fitting of the estimated dispersions.

For exploration of the data we performed a Principal component plot (PC plot). For PC1 vs. PC2, the parents are not as strongly clustered together as expected. The offspring lines are distributed over the whole plot, this is as expected. Still the plot PC3 vs. PC4, shows us that here there is clustering of the parents towards either side, with all lines distributed in between, these PCs explain most likely the genetic variation. The variation seen in PC1 vs. PC2 could be due to segregation of loci spontaneous variation can occur, causing this distribution. This is for instance seen in transgressive segregation, in which the offspring lines have more variation that the parents. This could be due to recombination, as due to segregation of loci expression can change (Rieseberg et al., 1999). Possibly environmental factors could be involved, however unlikely due to our experimental design. Another possibility is that there are technical errors affecting the data, possibly from pre-processing the data.

Differential expression analysis parents

The methods DESeq2, EdgeR and Ballgown were compared for differential expression analysis between the parents. There is more overlap in differentially expressed genes identified with DESeq2 and EdgeR than with Ballgown. This is most probably because DESeq2 and Edger have very similar methods with a similar approach. One difference is that Ballgown is working on a log2(1+FPKM) scale, while EdgeR and DESeq2 work with raw read counts. In addition, Ballgown uses an F-statistic while the other two use a generalized linear model fitting. Because of the big difference we chose to only work

further with genes identified with both EdgeR and DESeq2. Quite a lot of differentially expressed genes were identified. The parents, although both wheat, are relatively distant varieties. 200 years of intensive breeding has caused a lot of trait changes and improvements (such as disease resistances, dwarfing etc.) in wheat, explaining the high number of differentially expressed genes.

Among the differentially expressed genes we identified enriched GO terms related to defense and stress response. This can be explained by the fact that during the course of wheat improvement a lot of disease resistance genes have been introgressed, and are thus only recently introduced in modern wheat. Resistances or tolerance to abiotic stresses are another important factor that has been selected for in wheat. We identified transcription factors by comparing the *A. thaliana* homologs found in our differentially expressed gene list to an *A. thaliana* transcription factor database. A lot of identified transcription factors were related to biological processes that include abiotic defense response. For instance, lot of genes related to the stress response hormone abscisic acid were identified. Other transcription factors included involvement in leaf development, such as *ARF3* and *KANADIs* (Eshed et al., 2005). This could be due that RNA was extracted from leaf tissue. Small difference in leaf morphology between the two species could be the cause of this. However, also other terms were identified such as involvement in the plant hormone Gibberellin (GA), which plays an important role in dwarfing. GA is involved in the elongation of a plant, whereby dwarfing is caused by GA-insensitivity. Two known selection genes were identified, one involved in dwarfing, i.e. the GA-insensitive gene *RHT-B1 (*Sourdille et al., 1998). The other known selection gene encodes the PPD-D1c protein. This protein is related to later flowering (Bentley et al., 2013). This gene is upregulated in the later flowering Red Fife in comparison to Stettler, and could thus possibly contribute to the flower time trait.

Genes from secondary metabolite pathways, predicted by PlantiSmash, were only for a low percentage represented in the differentially expressed gene list. Thus, no full pathways, but only few genes of a pathway were identified. This could be due to chance, as PlantiSmash identified in total 2987 genes, and we initially started with around 100.000 genes, so one would expect that at least a few genes are in here at random. In 6000 genes you could therefore expect around 180 genes by chance ($\frac{3000}{100.000} \times 6000$), this is very close to the actual number of identified PlantiSmash genes, i.e. 189. A lot of pathway components were identified by BlastKoala, also including various KEGG identifiers per pathway. Components in photosystem I and II could be involved in energy transduction, a possible explanation is due to adaptation of circadian clock. A study identified that when the circadian clock of *A. thaliana* matches to photoperiod of the surrounding environment, photosynthesis conversion becomes significantly more efficient, fixing more carbon and growing faster (Dodd et al., 2005). Several components of hormone signal transduction were identified. Two genes involved in Gibberellin suppression, which could be due to the dwarfing trait. This included the selection gene *RHT-B1*, which is identified in BlastKoala as a DELLA repressor. In line with our expectations, this gene was upregulated in Stettler, while the other identified gene, a repressor of this DELLA repressor, was upregulated in Red Fife. The modern variety Stettler is a semi-dwarf while Red Fife is non-dwarf, explaining this differential expression. Differential expression in abscisic acid downstream signaling components could involve changes in response to abiotic stress response.

Network analysis

To identify single regulatory factors that control the expression of many genes and to identify linked genes with cis-regulatory variation we used network analysis. As we expected, we identified modules that could be regulated by single regulatory factors. These were observed as modules without linked genes. We also identified modules with linked genes, i.e. most genes originated from a single chromosome, these modules are thought to be co-expressed due to variation in their upstream regions. In most cases modules with unlinked genes showed a strong positive correlation expression pattern, while modules with linked genes showed both positive and negative correlation. This could be explained by that modules with unlinked genes are regulated by single trans-acting factors, thus affecting many downstream genes located on different places along the genome. An explanation for the pattern observed in modules with linked genes is that these are co-expressed due segregation of mutations in their upstream regions. As our population is only F2 (and double haploid), the recombination frequency is relatively low. It is thus possible, that lines have either allele of one of the parents at a position, causing this both positive and negative correlation expression pattern among the lines. Both explanations should be confirmed by e-QTL mapping analysis, to exclude the effect of random factors or for instance chromatin remodeling, and as we do not have evidence for the genetic basis, i.e. we did not identify actual alleles causing the expression changes. In theory, environmental

factors could have influenced our results. However, we think environmental factors were excluded in our analysis, as our experimental design pooled samples from three different growth periods (two weeks apart), which where all placed at random every period.

To further test our hypothesis, we assessed biological processes related to modules with GO-analysis, using *A. thaliana* identifiers. Sometimes up to 6 wheat homologues genes were identified for a single Arabidopsis gene, this is due to wheat being a hexaploid, carrying six gene copies instead of two. It will be interesting for future research to investigate more the role of these homologs and how they evolved in relation to wheat improvement. We see a difference in significant GO-terms identified in module with linked or unlinked genes. Modules with unlinked genes show in most cases significant GO-terms, often related to response to external factors, such as defense response. In addition, photosynthesis and flower development are identified in two distinct modules. These are all terms expected to be involved in improvement of wheat.

The pink module showed a lot of GO-terms related to alternate splicing, epigenetics and flower development, and could thus possibly be involved in flowering time as Stettler is known to flower earlier than Red Fife. A known selection gene involved in flowering was identified, i.e. encoding PHYTOCHROME AND FLOWERING TIME PROTEIN 1. Interesting is that the transcriptional bias of the genes is towards the Red Fife parent, the heritage cultivar. Possibly this cultivar carries also early flowering alleles, or the genes in this module induce flowering time. The bias was assessed by looking which genes in the module were differentially expressed between the parents, and if these genes were all up or down regulated in one parent in comparison to the other. Although only few genes of the total module were represented in the differentially expressed gene list of the parent, one would expect that all genes are directed towards one parent if our hypothesis is true. This as if one single factor affects the expression of all downstream genes, these genes should overall be all upregulated or all downregulated, with exception for repressors. A previous study identified that upon temperature alterations in *A. thaliana*, the splicing machinery itself is alternatively spliced which in turn targets the alternative splicing of flower time genes (Verhage et al., 2017). Genes such as components of the circadian clock and flowering time genes, are sensitive to alternative splicing due to temperature fluctuations (Verhage et al., 2017). In other studies, the effect of alternative splicing on flowering time is stressed also out, for instance alternative splicing of the *flowering time control* (*FCA*) gene shows significant effect on promotion of flower development (Eckardt et al., 2002).

Linked genes co-expressed due to variation in their upstream regions were expected to be found among modules, however they are not per se biologically meaningful, as they can cluster together due to segregation by chance. We therefore question if these genetically linked genes are all upregulated or downregulated in the parents (biased) as this indicates selection and if they are functionally linked, i.e. involved in the same biological process. Only a small percentage of these modules showed significant enriched GO-terms, related to defense response. Whether these defense response enriched modules are selected for is doubtful, as defense response genes have been found to cluster together on the genome by previous research (Li et al., 1999, Clavijo et al., 2016). This could be the result of genome reorganizations, and does not necessarily mean it is due to selection. Also, testing whether genes of the module present in the differentially expressed gene list of the parents, had a bias towards one of the parents, did not identify a bias for any modules with significant GO-enrichment. Only two modules showed a bias towards one of the parents, but these did not include any significant GO-terms. However, it is not that surprising most of these linked cis-regulated modules do not show a bias towards one of the parents, as in the clustered heatmaps the parents were mostly located at either side, suggesting half of the lines would carry the alleles of either parent. In addition, we performed pathway analysis, no full pathways predicted by PlantiSmash were identified. For the modules we assessed with BlastKoala also no full pathways were identified, only very few KEGG identifiers per pathway. We thus did not find evidence that linked cis-factors are involved in the regulation of biological processes related to plant improvement in wheat, with maybe two exceptions as two modules showed a bias towards one of the parents.

In total 1452 differentially expressed genes of the parents were not found back in the modules. A possible explanation for this is that these genes are regulated by multiple unlinked cis-regulated loci. Due to segregation of these loci, these genes would not be co-expressed over the whole population and thus not found back in the network. However, this must be confirmed by e-QTL mapping analysis. Another way to verify our statements about variation in trans or cis-acting factors causing our results, is to look at the connectivity of the genes in the module. If in a module there are genes with high connectivity, it is an indication that these are important regulators (trans-acting factors) (Carlson et al., 2006). Alternatively,

genes within a module have only low connectivity, thus only few connections to other genes, could indicate that variation in upstream regions affects the regulation.

Conclusions

This thesis aimed to explore the involvement of gene regulatory variation in relation to wheat improvement, involving traits such as disease resistances, dwarfing and flowering time. Regulation due to multiple alleles is difficult to maintain in a segregating population. We therefore proposed that selection during wheat improvement would favor a simpler regulation, that includes either regulation by single master regulatory loci or by linked genes with variation in their upstream regions.

From our results, we conclude that single master regulatory loci that control the expression of many downstream genes play a role in the regulation for biological processes involved in wheat improvement. Several modules were identified that showed a strong positive correlation, and had low linkage of genes. For many of these modules we found evidence for involvement in biological processes related to wheat improvement, such as dwarfing and flowering time.

Even though we identified modules that are possibly correlated due to fixation of linked genes with cis-regulatory variation, we do not have strong evidence that these are involved in regulation for biological processes related to wheat improvement. This because only in two modules most genes were biased in the parents, i.e. all upregulated or downregulated towards one of the parents, which would indicate selection for this region. Although, these two modules were not enriched in GO-terms they can still potentially be biologically relevant. In addition, no full pathway components were found in modules with linked genes that could indicate functional linkage.

Still, a lot of differentially expressed genes between the parents were not identified in our modules, possibly due regulation by multiple unlinked cis-regulated loci. Overall, we state that for wheat improvement, the genetic control of many genes involves single regulatory genes for many cases. All our results should be combined with e-QTL analysis for validation, e.g. assess whether known selection genes are involved in genetic control, and to identify the genetic basis of variation. After this validation, possibly new genes can be identified that are the important regulators of genes involved in wheat improvement. These results may be applicable for other (complex polyploid) crops. This as selection for plant improvement might have caused similar gene regulatory variation for similar traits.

Literature

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59-60.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC genomics, 7(1), 40.

Chen, A., Li, C., Hu, W., Lau, M. Y., Lin, H., Rockwell, N. C., ... & Dubcovsky, J. (2014). PHYTOCHROME C plays a major role in the acceleration of wheat flowering under long-day photoperiod. *Proceedings of the National Academy of Sciences*, *111*(28), 10037-10044.

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., ... & Lipscombe, J. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome research, 27(5), 885-896.

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One*, *8*(12), e85024.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., ... & Webb, A. A. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. Science, 309(5734), 630-633.

Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006). The molecular genetics of crop domestication. Cell, 127(7), 1309-1321.

Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, *316*(5833), 1862-1866.

Eshed, Y., Baum, S. F., Perea, J. V., & Bowman, J. L. (2001). Establishment of polarity in lateral organs of plants. Current Biology, 11(16), 1251-1260.

Eckardt, N. A. (2002). Alternative splicing and the control of flowering time. The Plant cell, 14(4), 743-747.

Ford, R. H. (2000). Inheritance of kernel color in corn: explanations & investigations. The American Biology Teacher, 62(3), 181-188.

Hedden, P. (2003). The genes of the Green Revolution. *TRENDS in Genetics*, *19*(1), 5-9.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., ... & Tappu, R. (2016). MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS computational biology*, *12*(6), e1004957.

International Wheat Genome Sequencing Consortium (IWGSC). (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, *345*(6194), 1251788.

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology*, *428*(4), 726-731.

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic acids research*, *45*(W1), W55-W63.

Krueger, F. (2015). Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 559.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.

Meyer, R. S., & Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. Nature reviews genetics, 14(12), 840-852.

Pearce, S., Saville, R., Vaughan, S. P., Chandler, P. M., Wilhelm, E. P., Sparks, C. A., ... & Hedden, P. (2011). Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. *Plant physiology*, *157*(4), 1820-1831.

Peng, J., Ronin, Y., Fahima, T., Röder, M. S., Li, Y., Nevo, E., & Korol, A. (2003). Domestication quantitative trait loci in Triticum dicoccoides, the progenitor of wheat. *Proceedings of the National Academy of Sciences*, *100*(5), 2489-2494.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, *33*(3), 290-295.

Pont, C., & Salse, J. (2017). Wheat paleohistory created asymmetrical genomic evolution. *Current Opinion in Plant Biology*, *36*, 29-37.

Rieseberg, L. H., Archer, M. A., & Wayne, R. K. (1999). Transgressive segregation, adaptation and speciation. Heredity, 83(4), 363-372.

Sourdille, P., Charmet, G., Trottet, M., Tixier, M. H., Boeuf, C., Negre, S., ... & Bernard, M. (1998). Linkage between RFLP molecular markers and the dwarfing genes Rht-B1 and Rht-D1 in wheat. Hereditas, 128(1), 41-46.

Swinnen, G., Goossens, A., & Pauwels, L. (2016). Lessons from domestication: targeting cis-regulatory elements for crop improvement. Trends in plant science, 21(6), 506-515.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., ... & Su, Z. (2017). agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*.

Veeneman, B. A., Shukla, S., Dhanasekaran, S. M., Chinnaiyan, A. M., & Nesvizhskii, A. I. (2015). Two-pass alignment improves novel splice junction quantification. *Bioinformatics*, *32*(1), 43-49.

Verhage, L., Severing, E. I., Bucher, J., Lammers, M., Busscher-Lange, J., Bonnema, G., ... & Immink, R. G. (2017). Splicing-related genes are alternatively spliced upon changes in ambient temperatures in plants. PloS one, 12(3), e0172950.

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., ... & Dubcovsky, J. (2004). The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science*, *303*(5664), 1640-1644.

Yasuda S, Shimoyama H (1965) Analysis of internal factors influencing the heading time of wheat varieties. Ber Ohara Inst Landw Biol Okayama Univ 13:23–38

Appendix

Table 4. GO-analysis results with upregulated differentially expressed genes between parents, i.e. these are upregulated in Stettler.

GO term	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006952	defense response	221	732	8.10E-20	6.10E-16
GO:0009607	response to biotic stimulus	157	589	1.00E-10	2.00E-07
GO:0050896	response to stimulus	640	3408	5.80E-11	2.00E-07
GO:0006950	response to stress	407	1985	8.00E-11	2.00E-07
GO:0051707	response to other organism	155	581	1.30E-10	2.00E-07
GO:0043207	response to external biotic stimulus	155	583	1.60E-10	2.00E-07
GO:0006468	protein phosphorylation	150	559	1.90E-10	2.00E-07
GO:0051704	multi-organism process	193	808	1.70E-09	1.40E-06
GO:0098542	defense response to other organism	120	428	1.60E-09	1.40E-06
GO:0009617	response to bacterium	91	291	2.20E-09	1.60E-06
GO:0016310	phosphorylation	180	765	1.60E-08	1.10E-05
GO:0042742	defense response to bacterium	75	243	8.70E-08	5.50E-05
GO:0044550	secondary metabolite biosynthetic process	52	144	1.80E-07	0.0001
GO:0009605	response to external stimulus	178	792	2.90E-07	0.00015
GO:0044711	single-organism biosynthetic process	223	1057	5.60E-07	0.00028
GO:0019748	secondary metabolic process	68	235	2.30E-06	0.0011
GO:0044710	single-organism metabolic process	439	2410	5.10E-06	0.0021
GO:0044699	single-organism process	887	5316	5.10E-06	0.0021
GO:0002376	immune system process	69	253	1.10E-05	0.0042
GO:0008152	metabolic process	1057	6503	1.60E-05	0.0061
GO:0007166	cell surface receptor signaling pathway	46	149	2.60E-05	0.0094
GO:0009699	phenylpropanoid biosynthetic process	28	71	3.40E-05	0.012
GO:0006955	immune response	61	229	5.90E-05	0.019
GO:0045087	innate immune response	59	220	6.60E-05	0.021
GO:0006793	phosphorus metabolic process	213	1090	7.70E-05	0.023
GO:0006796	phosphate-containing compound metabolic process	211	1080	8.40E-05	0.024
GO:0009751	response to salicylic acid	36	114	0.00013	0.036

Table 5. GO-analysis results with downregulated differentially expressed genes between parents, i.e. these genes are downregulated in Stettler.

GO term	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0006468	protein phosphorylation	106	559	5.10E-06	0.017
GO:0006950	response to stress	301	1985	3.80E-06	0.017
GO:0006952	defense response	129	732	1.20E-05	0.026
GO:0016310	phosphorylation	133	765	1.60E-05	0.026
GO:0050896	response to stimulus	472	3408	3.00E-05	0.04

Table 6. Small selection of A. thaliana transcription factor homologs identified in the D.E. gene list.

TF	Gene description	Biological processes
AT1G69120	APETALA1	cell differentiation, floral meristem determinacy, flower development, maintenance of floral meristem identity, meristem structural organization, positive regulation of transcription from RNA polymerase II promoter, positive regulation of transcription, DNA-templated, regulation of transcription, DNA-templated, transcription, DNA-templated
AT1G80840	WRKY DNA-binding protein 40	defense response to bacterium, defense response to fungus, negative regulation of transcription, DNA-templated, regulation of defense response, regulation of defense response to virus by host, regulation of transcription, DNA-templated,lated, response to chitin, response to molecule of bacterial origin, response to salicylic acid, response to wounding, transcription, DNA-templated
AT2G01570	Repressor of GA, RGA1. Member of the VHIID/DELLA regulatory family.	gibberellic acid mediated signaling pathway, hyperosmotic salinity response, jasmonic acid mediated signaling pathway, meiotic cytokinesis, multicellular organism development, negative regulation of gibberellic acid mediated signaling pathway, negative regulation of seed germination, regulation of protein catabolic process, regulation of reactive oxygen species metabolic process, regulation of seed dormancy process, regulation of seed germination, regulation of transcription, DNA-templated, response to abscisic acid, response to ethylene, response to far red light, response to salt stress, salicylic acid mediated signaling pathway, transcription, DNA-templated
AT2G03500	EARLY FLOWERING MYB PROTEIN; EFM	flower development, gibberellic acid mediated signaling pathway, histone H3-K36 methylation, negative regulation of long-day photoperiodism, flowering, negative regulation of nucleic acid-templated transcription, regulation of transcription, DNA-templated, response to temperature stimulus, transcription, DNA-templated
AT2G33860	ARF3, Auxin response factor 3	abaxial cell fate specification, auxin metabolic process, auxin-activated signaling pathway, floral meristem determinacy, regulation of transcription, DNA-templated, response to auxin, transcription, DNA-templated, vegetative phase change
AT2G46870	NGA1, AP2/B3-like transcriptional factor family protein	flower development, leaf development, regulation of leaf morphogenesis, regulation of transcription, DNA-templated, transcription, DNA-templated
AT3G56400	WRKY DNA-binding protein 70	defense response to bacterium, defense response to fungus, induced systemic resistance, jasmonic acid mediated signaling pathway, negative regulation of leaf senescence, negative regulation of transcription, DNA-templated, regulation of defense response, regulation of transcription, DNA-templated, response to chitin, response to jasmonic acid, response to salicylic acid, systemic acquired resistance, salicylic acid mediated signaling pathway, transcription, DNA-templated
AT4G11880	AGAMOUS-like 14	flower development, maintenance of floral meristem identity, positive regulation of transcription from RNA polymerase II promoter, regulation of auxin polar transport, regulation of root meristem growth, transcription, DNA-templated, vegetative to reproductive phase transition of meristem
AT5G65640	beta HLH protein 93	gibberellin catabolic process, multicellular organism development, regulation of gibberellin biosynthetic process, regulation of transcription, DNA-templated, transcription, DNA-templated

Table 7. Differential expression of small selection of TFs in the parents Red Fife (heritage) and Stettler (modern).

TF	Gene description	log2FoldChange	Upregulated in	pvalue
AT1G69120	APETALA1	-4.732529683	Red Fife	3.25E-09
AT1G80840	WRKY DNA-binding protein 40	4.345259631	Stettler	1.16E-05
AT2G01570	GRAS family transcription factor family protein [repressor of GA]	0.529965077	Stettler	1.56E-05
AT2G03500	EARLY FLOWERING MYB PROTEIN; EFM	1.156057298	Stettler	0.000278221
AT2G33860	ARF3, Auxin response factor 3	1.5971067	Stettler	0.002788694
AT2G46870	NGA1, AP2/B3-like transcriptional factor family protein	0.669954117	Stettler	0.003358567
AT3G56400	WRKY DNA-binding protein 70	3.125871584	Stettler	3.58E-06
AT4G11880	AGAMOUS-like 14	-4.095415999	Red Fife	0.000450772

AT5G65640	beta HLH protein 93	-0.711493679	Red Fife	0.002381715

Table 8. Differential expression of small selection of identified genes by KEGG in the parents Red Fife (heritage) and Stettler (modern).

Protein	Wheat ID	baseMean	log2FoldChange	Upregulated in	pvalue
GID2	TraesCS3B01G068100	20.13114021	-2.673757238	Red Fife	1.94E-08
DELLA	TraesCS4B01G043100	1451.579684	0.529965077	Stettler	1.56E-05
PP2C	TraesCS1D01G449700	56.5644922	1.236323691	Stettler	1.48E-05
SnRK2	TraesCS2A01G566700	1096.112837	1.375036757	Stettler	7.59E-08
ABF	TraesCS6A01G333600	65.8710473	-0.729750764	Red Fife	0.000423

Figure 15. Sample clustering to detect outliers.

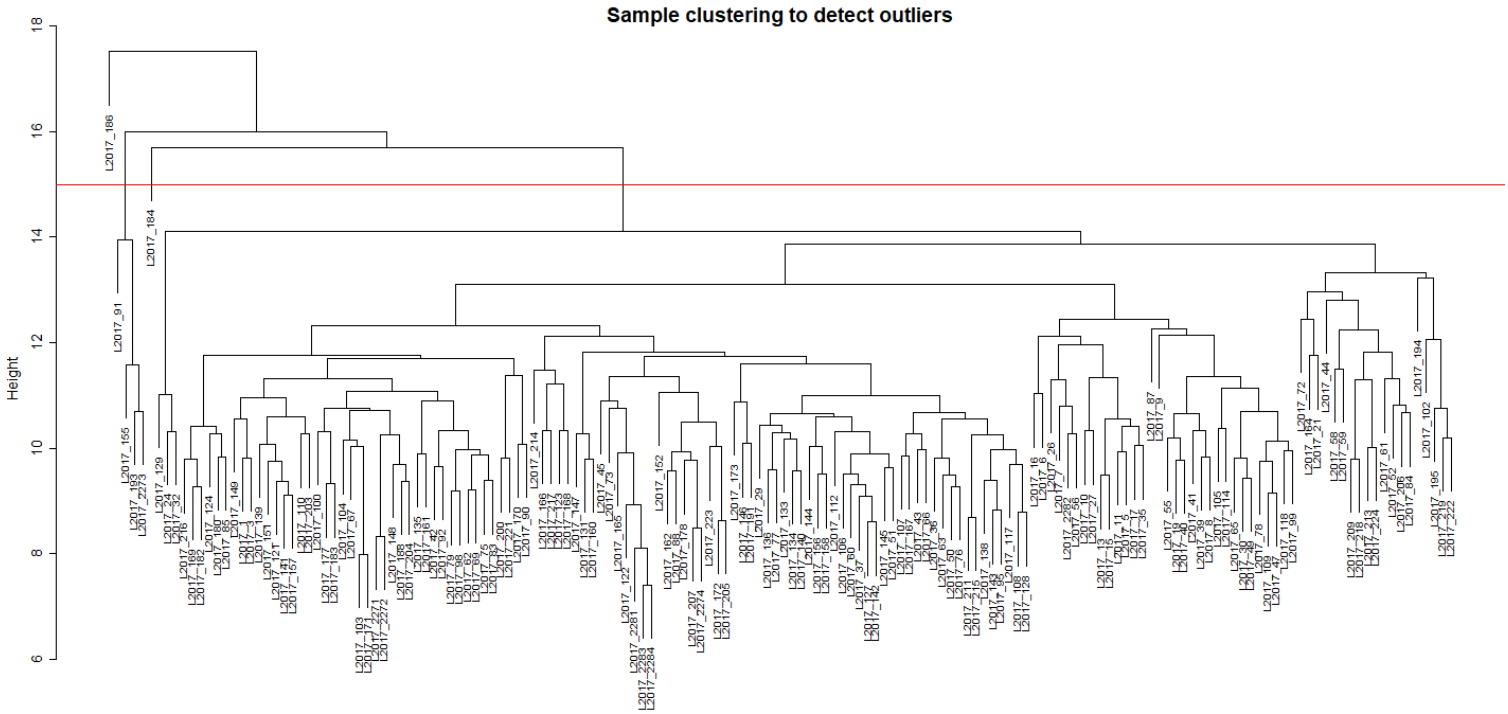


Figure 16. Module eigengene clustering

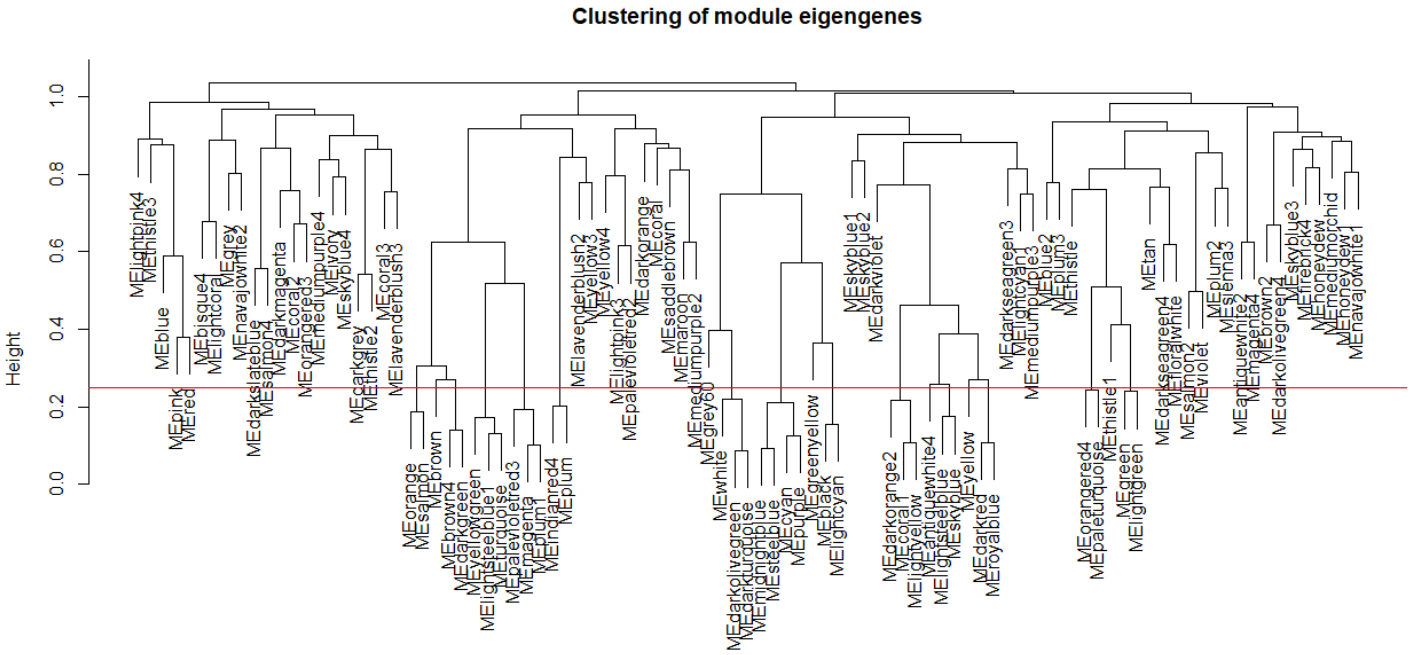


Table 9. Overview of module information. Includes module name, module size, correlation pattern (if only positive correlation over all lines then marked with +), the percentage of genes located on a single chromosome, the number of identified transcription factors (from blast results with *A. thaliana*), the number of known selection genes, the number of genes identified with plantismash and the percentage of *A. thaliana* identifiers obtained through blastx.

Module	Size	Correlation pattern	percentage genes on single chromosome	# Transcription factors [AT]	# Known genes	# Plantismash genes	percentage AT identifiers (blastx)
antiquewhite2	37		81.1	2	0	0	94.6
antiquewhite4	347	+	8.1	20	0	7	98.3
bisque4	111		90.1	6	0	1	78.4
black	3861		6.0	289	1	46	93.9
blue	2738		5.8	109	1	32	95.1
blue2	53		92.5	7	0	0	75.5
brown2	58		89.7	2	0	1	70.7
coral	37		75.7	0	0	0	89.2
coral2	75		61.3	7	2	2	94.7
coral3	34		94.1	1	0	2	64.7
darkgrey	278		86.7	8	0	3	87.8
darkmagenta	148		93.2	8	0	1	84.5
darkolivegreen4	58		65.5	6	0	0	84.5
darkorange	221	+	14.9	2	0	6	44.8
darkorange2	528	+	6.4	4	0	18	96.2
darkred	631	+	6.8	31	0	9	97.8
darkseagreen3	37		73.0	1	0	2	89.2
darkseagreen4	77	+	10.4	3	0	9	90.9
darkslateblue	105		64.8	3	0	4	92.4
darkviolet	51	+	37.3	1	0	0	88.2
firebrick4	58		81.0	0	0	4	81.0

floralwhite	128	+	7.8	4	0	6	96.1
green	1759	+	7.0	58	1	72	94.5
greenyellow	631	+	7.8	42	0	14	94.8
grey60	377	+	8.0	19	0	22	97.9
honeydew	38		94.7	4	0	18	89.5
honeydew1	83		89.2	2	1	1	90.4
indianred4	129		72.1	3	0	2	92.2
ivory	128		97.7	4	0	9	86.7
lavenderblush2	39		97.4	5	0	4	94.9
lavenderblush3	84		75.0	1	0	1	78.6
lightcoral	63		92.1	1	0	2	77.8
lightcyan1	128		87.5	5	0	0	82.8
lightpink3	39	+	12.8	0	0	2	87.2
lightpink4	85		89.4	7	1	4	84.7
magenta4	39		82.1	1	0	1	79.5
maroon	86	+	10.5	30	0	0	93.0
mediumorchid	74		95.9	3	0	3	87.8
mediumpurple2	66	+	13.6	2	0	2	89.4
mediumpurple3	130		60.0	6	0	3	85.4
mediumpurple4	33		97.0	1	0	3	84.8
midnightblue	1695		6.6	113	1	0	95.6
navajowhite1	39		100.0	1	0	37	87.2
navajowhite2	89		87.6	1	0	0	86.5
orange	726		7.2	25	0	11	96.7
orangered3	66		80.3	1	0	12	93.9
orangered4	313	+	8.3	0	0	4	98.1
palevioletred2	39	+	23.1	3	0	0	94.9
palevioletred3	5813		5.7	369	3	0	93.7
pink	1099	+	6.1	45	1	93	95.2
plum2	104		94.2	3	0	14	86.5
plum3	49		77.6	0	0	4	89.8
red	1259		6.8	96	0	0	93.7
saddlebrown	205		90.7	10	0	17	83.4
salmon2	43		76.7	0	0	2	90.7
salmon4	98		92.9	3	0	5	76.5
sienna3	144		56.3	5	0	0	88.2
skyblue1	68	+	88.2	3	0	1	88.2
skyblue2	72		56.9	1	0	0	95.8
skyblue3	143		90.9	4	0	2	81.1
skyblue4	32		87.5	2	0	0	84.4
tan	523	+	6.3	117	0	0	89.7
thistle	43		83.7	5	0	9	79.1
thistle1	98		8.2	1	0	3	88.8
thistle2	99		66.7	9	0	0	86.9
thistle3	45		93.3	1	0	1	95.6
violet	165		90.9	6	0	2	83.6
white	676	+	6.7	23	0	2	95.0
yellow	1494		6.6	60	0	25	93.6
yellow3	30		60.0	1	0	34	96.7
yellow4	72		81.9	1	0	2	88.9
Parents	6039			229	2	6	85.7
Parents up	3581						88.0
Parents down	2456						82.4

Table 10. Enriched gene ontology terms in the pink module

GO term	Description	Number in input list	Number in BG/Ref	p-value	FDR
GO:0051276	chromosome organization	60	217	2.40E-20	9.10E-17
GO:0006325	chromatin organization	46	168	1.10E-15	1.40E-12
GO:0090304	nucleic acid metabolic process	156	1355	1.10E-15	1.40E-12
GO:0010467	gene expression	162	1498	4.50E-14	4.30E-11
GO:0040029	regulation of gene expression, epigenetic	33	97	9.60E-14	7.30E-11
GO:0010629	negative regulation of gene expression	45	187	1.20E-13	7.70E-11
GO:0010605	negative regulation of macromolecule metabolic process	48	216	2.30E-13	1.30E-10
GO:0016569	covalent chromatin modification	36	123	3.10E-13	1.40E-10
GO:0016070	RNA metabolic process	138	1229	5.70E-13	2.40E-10
GO:0010468	regulation of gene expression	116	962	1.10E-12	4.00E-10
GO:0060255	regulation of macromolecule metabolic process	122	1042	1.50E-12	5.20E-10
GO:0006139	nucleobase-containing compound metabolic process	162	1569	1.80E-12	5.70E-10
GO:0016458	gene silencing	32	106	3.30E-12	9.70E-10
GO:0009892	negative regulation of metabolic process	48	240	6.40E-12	1.70E-09
GO:0019222	regulation of metabolic process	129	1157	6.80E-12	1.70E-09
GO:0006996	organelle organization	87	650	1.00E-11	2.40E-09
GO:0044260	cellular macromolecule metabolic process	240	2744	1.30E-11	2.80E-09
GO:0046483	heterocycle metabolic process	169	1728	3.40E-11	6.90E-09
GO:2000113	negative regulation of cellular macromolecule biosynthetic process	35	141	3.40E-11	6.90E-09
GO:0050789	regulation of biological process	190	2043	6.80E-11	1.30E-08
GO:0043170	macromolecule metabolic process	253	2993	7.40E-11	1.30E-08
GO:0051171	regulation of nitrogen compound metabolic process	108	936	9.20E-11	1.60E-08
GO:0043933	macromolecular complex subunit organization	61	392	9.70E-11	1.60E-08
GO:0016071	mRNA metabolic process	43	217	1.10E-10	1.70E-08
GO:0010558	negative regulation of macromolecule biosynthetic process	35	148	1.10E-10	1.70E-08
GO:0051172	negative regulation of nitrogen compound metabolic process	36	158	1.40E-10	2.10E-08
GO:0031323	regulation of cellular metabolic process	117	1063	1.90E-10	2.60E-08
GO:0080090	regulation of primary metabolic process	112	1006	2.80E-10	3.70E-08
GO:0048608	reproductive structure development	76	575	4.30E-10	5.40E-08
GO:0061458	reproductive system development	76	575	4.30E-10	5.40E-08
GO:0006396	RNA processing	53	327	5.00E-10	5.90E-08
GO:0031327	negative regulation of cellular biosynthetic process	35	158	5.00E-10	5.90E-08
GO:2000112	regulation of cellular macromolecule biosynthetic process	102	891	5.20E-10	6.00E-08
GO:0010556	regulation of macromolecule biosynthetic process	102	894	6.20E-10	6.90E-08
GO:0009890	negative regulation of biosynthetic process	35	161	7.70E-10	8.00E-08
GO:0031324	negative regulation of cellular metabolic process	38	187	7.50E-10	8.00E-08
GO:0006725	cellular aromatic compound metabolic process	168	1794	1.00E-09	1.10E-07
GO:0003006	developmental process involved in reproduction	82	659	1.10E-09	1.10E-07
GO:0065007	biological regulation	197	2223	1.40E-09	1.30E-07
GO:0031326	regulation of cellular biosynthetic process	104	933	1.30E-09	1.30E-07
GO:0050794	regulation of cellular process	169	1826	2.10E-09	1.90E-07
GO:0009889	regulation of biosynthetic process	104	947	2.80E-09	2.50E-07
GO:0018205	peptidyl-lysine modification	22	68	2.90E-09	2.60E-07
GO:0097659	nucleic acid-templated transcription	96	849	3.40E-09	2.90E-07
GO:0006351	transcription, DNA-templated	96	849	3.40E-09	2.90E-07
GO:0022414	reproductive process	87	738	3.60E-09	2.90E-07
GO:0032774	RNA biosynthetic process	96	851	3.80E-09	3.10E-07
GO:0016043	cellular component organization	116	1110	4.00E-09	3.10E-07
GO:0006397	mRNA processing	36	182	4.00E-09	3.10E-07
GO:0048519	negative regulation of biological process	57	393	4.30E-09	3.30E-07
GO:0034641	cellular nitrogen compound metabolic process	180	2005	4.40E-09	3.30E-07
GO:0009790	embryo development	46	279	4.70E-09	3.40E-07
GO:0000003	reproduction	87	743	4.80E-09	3.40E-07

GO:0048523	negative regulation of cellular process	44	261	5.70E-09	4.00E-07
GO:1901360	organic cyclic compound metabolic process	169	1855	6.40E-09	4.40E-07
GO:0019219	regulation of nucleobase-containing compound metabolic process	95	853	8.50E-09	5.70E-07
GO:0016570	histone modification	23	83	1.50E-08	1.00E-06
GO:0051252	regulation of RNA metabolic process	92	828	1.70E-08	1.10E-06
GO:0009791	post-embryonic development	83	718	2.00E-08	1.30E-06
GO:2001141	regulation of RNA biosynthetic process	90	807	2.10E-08	1.30E-06
GO:0006355	regulation of transcription, DNA-templated	90	807	2.10E-08	1.30E-06
GO:1903506	regulation of nucleic acid-templated transcription	90	807	2.10E-08	1.30E-06
GO:0007389	pattern specification process	21	73	3.70E-08	2.20E-06
GO:0031047	gene silencing by RNA	20	68	5.80E-08	3.40E-06
GO:0071840	cellular component organization or biogenesis	120	1224	6.20E-08	3.60E-06
GO:0009893	positive regulation of metabolic process	36	209	9.10E-08	5.20E-06
GO:0044702	single organism reproductive process	76	659	9.30E-08	5.30E-06
GO:0048731	system development	91	853	1.20E-07	6.60E-06
GO:0010604	positive regulation of macromolecule metabolic process	32	174	1.30E-07	7.10E-06
GO:0051128	regulation of cellular component organization	29	148	1.60E-07	8.60E-06
GO:0034654	nucleobase-containing compound biosynthetic process	98	953	1.80E-07	9.60E-06
GO:0006807	nitrogen compound metabolic process	188	2239	2.20E-07	1.20E-05
GO:0045892	negative regulation of transcription, DNA-templated	24	107	2.30E-07	1.20E-05
GO:0016441	posttranscriptional gene silencing	16	47	2.40E-07	1.20E-05
GO:0034645	cellular macromolecule biosynthetic process	127	1366	3.80E-07	1.90E-05
GO:0003002	regionalization	18	64	4.70E-07	2.30E-05
GO:0048518	positive regulation of biological process	50	377	4.90E-07	2.40E-05
GO:0045934	negative regulation of nucleobase-containing compound metabolic process	25	122	5.60E-07	2.70E-05
GO:0009793	embryo development ending in seed dormancy	36	228	6.00E-07	2.90E-05
GO:1903507	negative regulation of nucleic acid-templated transcription	24	114	6.10E-07	2.90E-05
GO:1902679	negative regulation of RNA biosynthetic process	24	114	6.10E-07	2.90E-05
GO:0008380	RNA splicing	26	132	6.40E-07	3.00E-05
GO:0051253	negative regulation of RNA metabolic process	24	115	7.10E-07	3.20E-05
GO:0009059	macromolecule biosynthetic process	127	1389	8.80E-07	4.00E-05
GO:0006259	DNA metabolic process	33	202	9.10E-07	4.00E-05
GO:0018130	heterocycle biosynthetic process	104	1076	1.10E-06	4.90E-05
GO:0048522	positive regulation of cellular process	39	268	1.30E-06	5.50E-05
GO:0031325	positive regulation of cellular metabolic process	31	187	1.50E-06	6.40E-05
GO:0010608	posttranscriptional regulation of gene expression	23	112	1.50E-06	6.50E-05
GO:0032268	regulation of cellular protein metabolic process	26	144	2.70E-06	0.00011
GO:0035194	posttranscriptional gene silencing by RNA	13	38	3.20E-06	0.00013
GO:0033044	regulation of chromosome organization	13	38	3.20E-06	0.00013
GO:0010228	vegetative to reproductive phase transition of meristem	21	100	3.20E-06	0.00013
GO:0043414	macromolecule methylation	18	76	3.80E-06	0.00015
GO:0006342	chromatin silencing	13	39	4.00E-06	0.00016
GO:0048316	seed development	40	295	4.40E-06	0.00017
GO:0090567	reproductive shoot system development	35	241	4.70E-06	0.00018
GO:0045814	negative regulation of gene expression, epigenetic	13	41	6.30E-06	0.00024
GO:0010154	fruit development	41	312	6.70E-06	0.00026
GO:0048856	anatomical structure development	119	1341	8.60E-06	0.00032
GO:0000819	sister chromatid segregation	9	18	9.80E-06	0.00037
GO:0010628	positive regulation of gene expression	24	137	1.00E-05	0.00038
GO:0007059	chromosome segregation	12	37	1.20E-05	0.00043
GO:0007275	multicellular organism development	108	1196	1.20E-05	0.00045
GO:0034968	histone lysine methylation	11	31	1.40E-05	0.00049
GO:0019438	aromatic compound biosynthetic process	102	1115	1.40E-05	0.00049
GO:0009908	flower development	33	233	1.40E-05	0.00049
GO:0051246	regulation of protein metabolic process	26	160	1.40E-05	0.0005
GO:0017148	negative regulation of translation	11	32	1.80E-05	0.00061

GO:0044707	single-multicellular organism process	109	1223	1.90E-05	0.00066
GO:0032502	developmental process	120	1383	2.10E-05	0.00071
GO:0044767	single-organism developmental process	117	1341	2.20E-05	0.00073
GO:0018022	peptidyl-lysine methylation	11	33	2.20E-05	0.00074
GO:0034249	negative regulation of cellular amide metabolic process	11	33	2.20E-05	0.00074
GO:0050793	regulation of developmental process	37	284	2.30E-05	0.00074
GO:0051248	negative regulation of protein metabolic process	13	48	2.60E-05	0.00084
GO:0032269	negative regulation of cellular protein metabolic process	13	48	2.60E-05	0.00084
GO:0035195	gene silencing by miRNA	9	22	3.40E-05	0.0011
GO:1901362	organic cyclic compound biosynthetic process	104	1178	4.50E-05	0.0014
GO:0043412	macromolecule modification	103	1165	4.70E-05	0.0015
GO:0044271	cellular nitrogen compound biosynthetic process	120	1411	4.90E-05	0.0015
GO:0032259	methylation	22	133	5.10E-05	0.0016
GO:0032501	multicellular organismal process	111	1283	5.20E-05	0.0016
GO:0048367	shoot system development	45	394	5.80E-05	0.0018
GO:0016571	histone methylation	11	38	6.50E-05	0.002
GO:0098813	nuclear chromosome segregation	10	31	6.60E-05	0.002
GO:1902275	regulation of chromatin organization	10	31	6.60E-05	0.002
GO:0016573	histone acetylation	9	25	7.70E-05	0.0022
GO:0018393	internal peptidyl-lysine acetylation	9	25	7.70E-05	0.0022
GO:0018394	peptidyl-lysine acetylation	9	25	7.70E-05	0.0022
GO:0033043	regulation of organelle organization	16	80	7.90E-05	0.0023
GO:0044238	primary metabolic process	274	3841	8.40E-05	0.0024
GO:0009987	cellular process	346	5059	9.10E-05	0.0026
GO:0006366	transcription from RNA polymerase II promoter	18	100	9.50E-05	0.0027
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	21	131	0.00011	0.0031
GO:0006475	internal protein amino acid acetylation	9	27	0.00012	0.0035
GO:2000026	regulation of multicellular organismal development	29	218	0.00013	0.0035
GO:0018193	peptidyl-amino acid modification	23	153	0.00013	0.0035
GO:0008213	protein alkylation	11	42	0.00014	0.0038
GO:0006479	protein methylation	11	42	0.00014	0.0038
GO:0060968	regulation of gene silencing	8	21	0.00014	0.0038
GO:0051173	positive regulation of nitrogen compound metabolic process	22	144	0.00014	0.0038
GO:0007033	vacuole organization	9	28	0.00016	0.0041
GO:0006473	protein acetylation	9	29	0.0002	0.0051
GO:0010557	positive regulation of macromolecule biosynthetic process	20	127	0.0002	0.0051
GO:0006974	cellular response to DNA damage stimulus	20	127	0.0002	0.0051
GO:0051239	regulation of multicellular organismal process	30	236	0.0002	0.0051
GO:0006304	DNA modification	10	37	0.00022	0.0057
GO:0009891	positive regulation of biosynthetic process	22	149	0.00022	0.0057
GO:0044728	DNA methylation or demethylation	10	37	0.00022	0.0057
GO:0009909	regulation of flower development	15	80	0.00024	0.0061
GO:0044237	cellular metabolic process	280	4002	0.00028	0.0069
GO:0031328	positive regulation of cellular biosynthetic process	21	141	0.00028	0.0069
GO:0006417	regulation of translation	15	82	0.00031	0.0075
GO:0006378	mRNA polyadenylation	6	12	0.00032	0.0076
GO:0043631	RNA polyadenylation	6	12	0.00032	0.0076
GO:0007062	sister chromatid cohesion	6	12	0.00032	0.0076
GO:0034248	regulation of cellular amide metabolic process	15	83	0.00034	0.0082
GO:0051254	positive regulation of RNA metabolic process	19	124	0.00039	0.0093
GO:0006306	DNA methylation	9	33	0.00044	0.01
GO:0006305	DNA alkylation	9	33	0.00044	0.01
GO:0006338	chromatin remodeling	7	19	0.00044	0.01
GO:0000956	nuclear-transcribed mRNA catabolic process	9	33	0.00044	0.01
GO:0035196	production of miRNAs involved in gene silencing by miRNA	6	13	0.00044	0.01
GO:0048831	regulation of shoot system development	17	105	0.00045	0.01

GO:0006357	regulation of transcription from RNA polymerase II promoter	14	76	0.00045	0.01
GO:0007067	mitotic nuclear division	10	41	0.00045	0.01
GO:2000241	regulation of reproductive process	19	126	0.00047	0.01
GO:0071103	DNA conformation change	8	26	0.00046	0.01
GO:1902680	positive regulation of RNA biosynthetic process	18	116	0.00048	0.011
GO:0045893	positive regulation of transcription, DNA-templated	18	116	0.00048	0.011
GO:1903508	positive regulation of nucleic acid-templated transcription	18	116	0.00048	0.011
GO:0061647	histone H3-K9 modification	6	14	0.0006	0.013
GO:0000280	nuclear division	13	69	0.0006	0.013
GO:0031124	mRNA 3'-end processing	6	14	0.0006	0.013
GO:0043543	protein acylation	11	51	0.00059	0.013
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	6	14	0.0006	0.013
GO:0043331	response to dsRNA	8	28	0.0007	0.015
GO:0031050	dsRNA fragmentation	8	28	0.0007	0.015
GO:0070918	production of small RNA involved in gene silencing by RNA	8	28	0.0007	0.015
GO:0071359	cellular response to dsRNA	8	28	0.0007	0.015
GO:0033554	cellular response to stress	42	408	0.00076	0.016
GO:0031399	regulation of protein modification process	11	53	0.00078	0.016
GO:0031060	regulation of histone methylation	6	15	0.0008	0.016
GO:0048580	regulation of post-embryonic development	22	166	0.00084	0.017
GO:0006401	RNA catabolic process	10	45	0.00085	0.017
GO:0006402	mRNA catabolic process	9	37	0.00088	0.018
GO:0031056	regulation of histone modification	7	22	0.00089	0.018
GO:0006281	DNA repair	17	114	0.001	0.02
GO:0031123	RNA 3'-end processing	7	23	0.0011	0.022
GO:0000398	mRNA splicing, via spliceosome	12	65	0.0011	0.022
GO:1901699	cellular response to nitrogen compound	10	47	0.0011	0.022
GO:0071704	organic substance metabolic process	288	4226	0.0013	0.026
GO:0031935	regulation of chromatin silencing	5	11	0.0014	0.028
GO:0006914	autophagy	8	32	0.0015	0.028
GO:1902589	single-organism organelle organization	19	141	0.0016	0.03
GO:0022402	cell cycle process	18	131	0.0017	0.032
GO:0007049	cell cycle	23	190	0.0019	0.037
GO:0000070	mitotic sister chromatid segregation	5	12	0.0019	0.037
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	7	26	0.002	0.038
GO:0044249	cellular biosynthetic process	155	2097	0.002	0.038
GO:0051130	positive regulation of cellular component organization	9	43	0.0022	0.041
GO:0051567	histone H3-K9 methylation	5	13	0.0025	0.047
GO:0006333	chromatin assembly or disassembly	6	20	0.0027	0.049
GO:0000375	RNA splicing, via transesterification reactions	12	73	0.0027	0.049
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	12	73	0.0027	0.049

Table 11. Small selection of identified transcription factors in pink module.

TF	Description	Biological process
AT1G43850	transcriptional co-regulator of AGAMOUS, that functions with LEUNIG to repress AG in the outer floral whorls	cell differentiation, cellular response to DNA damage stimulus, cellular response to external biotic stimulus, embryo development, gynoecium development,multicellular organism development, plant ovule development, regulation of flower development, regulation of transcription, DNA-templated, response to auxin,response to bacterium, response to cycloheximide, response to fungus, response to hypoxia, response to nematode, response to oxidative stress, response to silver ion, transcription, DNA-templated
AT2G23740	SU(VAR)3-9-RELATED PROTEIN 5	chromatin silencing, regulation of histone H3-K9 dimethylation

AT4G02560	LD, nuclear localized protein with similarity to transcriptional regulators.	cell differentiation, flower development, positive regulation of flower development,regulation of transcription, DNA-templated, vegetative to reproductive phase transition of meristem
AT4G04885	PCFS4 (Pcf11p-similar protein 4)	flower development, mRNA cleavage, mRNA polyadenylation, mRNA processing,positive regulation of flower development, termination of RNA polymerase II transcription
AT5G20730	Auxin response factor 7	auxin-activated signaling pathway, blue light signaling pathway, gravitropism,lateral root development, lateral root formation, leaf development, phototropism,positive regulation of transcription, DNA-templated, regulation of growth,regulation of transcription, DNA-templated, response to auxin, response to ethylene, transcription, DNA-templated
AT5G44180	Homeodomain-like transcriptional regulator	flower development, negative regulation of transcription, DNA-templated,regulation of transcription, DNA-templated, transcription, DNA-templated,vegetative to reproductive phase transition of meristem

Table 12. Small selection of identified transcription factors in green module.

TF	Description	Biological process
AT5G26170	WRKY DNA-binding protein 50	defense response to fungus, jasmonic acid mediated signaling pathway, regulation of transcription, DNA-templated, transcription, DNA-templated
AT4G17750	Heat shock factor 1	response to heat, regulation of transcription, DNA-templated, transcription, DNA-templated
AT3G56400	WRKY DNA-binding protein 70	defense response to bacterium, defense response to fungus, induced systemic resistance, jasmonic acid mediated signaling pathway, negative regulation of leaf senescence, negative regulation of transcription, DNA-templated, regulation of defense response, regulation of transcription, DNA-templated, response to chitin, response to jasmonic acid, response to salicylic acid, systemic acquired resistance, salicylic acid mediated signaling pathway, transcription, DNA-templated
AT2G40950	BZIP17	hyperosmotic salinity response, positive regulation of transcription, DNA-templated, regulation of transcription, DNA-templated, transcription, DNA-templated
AT1G53910	Member of the ERF (ethylene response factor) subfamily B-2 of ERF/AP2 transcription factor family	detection of hypoxia, ethylene-activated signaling pathway, regulation of root development, regulation of transcription, DNA-templated, response to hypoxia, transcription, DNA-templated

Table 13. All identified transcription factors in module blue2, including description and involvement in biological processes.

TF	Description	Biological process
AT1G71930	vascular related NAC-domain protein 7	cell wall organization, cellular response to auxin stimulus, defense response to fungus, multicellular organism development, positive regulation of gene expression, positive regulation of transcription, DNA-templated, protoxylem development, regulation of transcription, DNA-templated, response to abscisic acid, response to auxin, response to brassinosteroid, response to cytokinin, response to fungus, transcription, DNA-templated, xylan metabolic process, xylem development, xylem vessel member cell differentiation
AT1G75710	C2H2-like zinc finger protein	NA
AT4G36160	NAC domain containing protein 76	cell wall organization, multicellular organism development, positive regulation of secondary cell wall biogenesis, regulation of transcription, DNA-templated, transcription, DNA-templated, xylem vessel member cell differentiation
AT4G36920	AP2, Integrase-type DNA-binding superfamily protein	cell differentiation, flower development, meristem maintenance, plant ovule development, regulation of transcription, DNA-templated, seed development, sexual reproduction, specification of floral organ identity, transcription, DNA-templated